Methods for Monitoring Process Control and Capability in the Presence of Autocorrelation

DISSERTATION

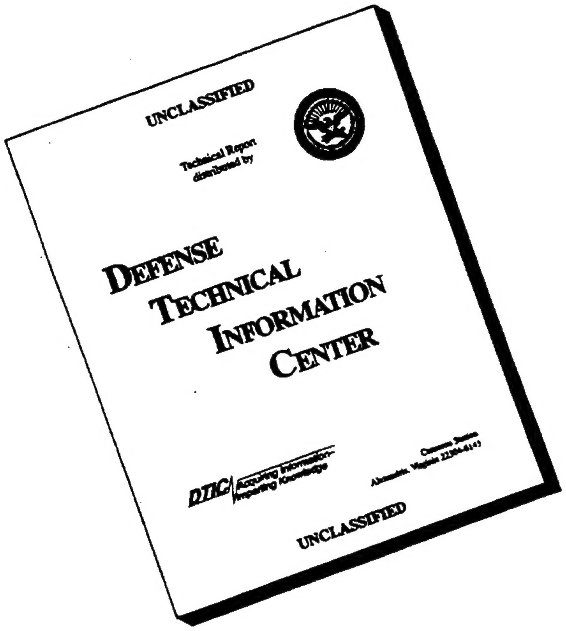
Daniel J. Zalewski
Major, USAF

AFIT/DS/ENS/95-02

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY
AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISCLAIMER NOTICE



UNCLASSIFIED

Technical Report distributed by

DEFENSE TECHNICAL INFORMATION CENTER

DTIC Acquiring Information— Imparting Knowledge

Cameron Station
Alexandria, Virginia 22304-6145

UNCLASSIFIED

THIS DOCUMENT IS BEST QUALITY AVAILABLE. THE COPY FURNISHED TO DTIC CONTAINED A SIGNIFICANT NUMBER OF PAGES WHICH DO NOT REPRODUCE LEGIBLY.

AFIT/DS/ENS/95-02

Methods for Monitoring Process Control and Capability in the Presence of Autocorrelation

DISSERTATION

Daniel J. Zalewski
Major, USAF

AFIT/DS/ENS/95-02

AFIT/DS/ENS/95-02

Methods for Monitoring Process Control and Capability in the Presence of Autocorrelation

Daniel J. Zalewski, B.S., M.S.
Major, USAF

Approved:

Edward F. Mykytka — 28 Aug 95

Dr Edward F. Mykytka, Chairman — Date

Paul F Auclair — 28 Aug 95

Lt Col Paul F. Auclair, Committee Member — Date

Kenneth W Bauer Jr — 28 AUG 95

Lt Col Kenneth W. Bauer, Jr., Committee Member — Date

Daniel E. Reynolds — 28 Aug 95

Prof Daniel E. Reynolds, Committee Member — Date

Albert H Moore — 29 Aug 95

Dr Albert H. Moore, Dean's Representative — Date

Accepted:

29 August 1995

Dr. Robert A. Calico, Jr., Dean — Date

Preface

The goal of this research was to develop a method for quality improvement based upon monitoring process capability. The feasibility and value of doing so are shown in this document.

My success in achieving this goal is the result of the assistance, guidance, and support of many people. First off, I would like to thank the United States Air Force for sponsoring me in the pursuit of this research. I am most indebted to my committee chairman, Dr Ed Mykytka, who kept me from swerving off into left field many times. I also want to thank the other members of my committee: the Dean's representative, Dr Al Moore; Lt Col Ken Bauer for his patience and questions reviewing the mathematics; Lt Col Paul Auclair for forcing me to hone my presentation skills; and especially Prof Dan Reynolds for his infectious enthusiasm and optimism.

Equally valued was the friendship and comraderie of my peers. Thanks to the first two who showed it was possible to finish: Mark Gallagher and Jean Steppe. Also, thanks goes to my officemate, Dennis Benson, who helped me keep things in perspective. And special thanks to Lisa Belue, who lead me through the program from start to finish. Without her showing me the way, I would probably have gotten lost.

Most important of all is the support of my family. I would like to thank my wife Julie for her encouragement and faith in me. Finally, I would like to thank my daughter Zoie whose arrival gave me the motivation I needed to push through to the end of my program. The largest price I paid in finishing this dissertation was missing the first time she rolled over.

Daniel J. Zalewski

Table of Contents

Page

Page

List of Figures

List of Tables

AFIT/DS/ENS/95-02

Abstract

When standard control charts are applied to a process whose measurements of quality exhibit autocorrelation, the performance of those charts can be considerably different than that expected when no autocorrelation is present. To model this performance, we extend the existing definitions of assignable and chance causes of variation to account for the variation induced by the autocorrelation structure. The application of statistical thinking toward continuous process improvement is discussed using the proposed taxonomy. We develop a method, suitable for use on processes whose behavior can be modelled as an ARMA(1,1) process, to select control limits which yield a specified average run length in the absence of assignable causes of variation.

The current paradigm for process improvement is centered around monitoring the state of statistical control. A new paradigm, based upon monitoring process capability instead of statistical control, is proposed. We differentiate between time-specific, time-average, and long-term capability, and propose both probability-based and loss-based measures of capability.

A capability monitoring system is developed for stationary ARMA(1,1) processes. This system is tested by simulating its response under a variety of mean shifts. The added value of knowing the parameters of the ARMA(1,1) process is also discussed. The benefits of this additional knowledge are demonstrated by simulating the response of capability monitoring systems tailored to independent normal and mixed ARMA(1,1) models to shifts in the mean and variance. The results are compared to previously published results for other methods.

Methods for Monitoring
Process Control and Capability
in the Presence of Autocorrelation

I. Introduction

The birth of Statistical Process Control (SPC) can be traced to Shewhart's 1931 book, "Economic Control of Quality of Manufactured Product." In that seminal work, Shewhart set forth principles for monitoring a process which remain in use today. In the days between World War I and World War II, manufacturing processes were relatively slow and were plagued by inconsistent product quality. Shewhart delivered functional tools that enabled industry to gain a measure of understanding and control over these processes. Shewhart's principles and tools have led, over time, to a wide range of methods for monitoring and controlling processes that are integral to the quality revolution in industry today.

Since World War II, two gradual changes in manufacturing processes have impacted upon the utility of strict application of the Shewhart control charts. First, production rates have increased, and second, automatic sensing devices to measure and record product quality characteristics have become prevalent. While quality measurements in the past could often be adequately modeled as independent and identically distributed, these two changes have produced more situations in which observations exhibit autocorrelation, particularly in continuous-process industries such as chemicals, pulp and paper, and mineral processing (MacGregor and Harris, 1990). Numerous other examples of autocorrelated processes can be found in Montgomery (1991) and Box and Jenkins (1976). Although Shewhart control charts were developed under the assumption of independence, they have proven to be robust to small degrees of dependence in the observed quality measurements. However, the high level of dependence routinely found in many processes today often results in unpredictable performance by a Shewhart control chart and incorrect inferences about the process. These inferences may, in turn, result in a decrease in the level of understanding and control over the process.

Even when a suitable control chart is used and the monitored process is found to be operating in a state of statistical control, the products produced by that process may not be suitable for their intended purpose. The state of statistical control is related to the capability of a process, but is not the only factor determining capability. A significant amount of recent work we discuss in Chapter II centers on measuring the capability of an in-control process. A natural question we raise in this research is whether process capability can be monitored directly in lieu of assessing capability indirectly by monitoring process control.

In this research, we attempt to take another step in the evolution of Statistical Process Control by developing a practical method for monitoring the capability of processes that generate autocorrelated observations. To do so, we develop three major themes. First, we develop a method for determining fixed control limits with known performance characteristics for a low order autoregressive moving average model. Second, we present the theoretical background and motivation for changing the focus from monitoring the state of statistical control of a process to monitoring the capability of a process. Third, we develop and test a practical method for monitoring process capability.

1.1 Dissertation Overview

This dissertation is organized into six chapters. This chapter includes a brief introduction and a set of basic definitions. The next chapter provides detailed background information about the SPC techniques in use today and their associated limitations. Chapter III presents a method for determining appropriate control limits for a known low order autoregressive moving average model. Chapter IV then develops the mathematical foundation for monitoring process capability, and Chapter V proposes a methodology for the practical implementation of a capability monitoring system. Chapter VI contains the conclusions of this research and recommendations for future work.

1.2 Basic Definitions

In this section, we define the basic concepts used in this research, starting with the definition of a process. Our examination of the concept of process quality leads to an exploration of process variation. We define statistical process control and establish its relationship to statistical thinking. We also explore the emphasis placed on the 'state of statistical control' in the literature. Finally, we examine the concept of process capability.

1.2.1 Process and Process Measurements. Pritsker (1986) defines a process as "a time-ordered sequence of events (which) may encompass several activities." Beauregard (1992) similarly defines a process as "a sequence of events with an input, value-added, and an output." While not mathematically rigorous, these two definitions of a process provide the basic level of understanding germane to this research.

We make the implicit assumption that some aspect of a process related to the quality of its output can be observed and measured over time. For example, in a parts production process the diameter of a hole drilled in a wing rib may be of critical importance to the final product (Montgomery and Friedman, 1989). In this example, measurements are made on every hole drilled and are recorded in the time order of production. Inferences about the underlying process are then tested by analyzing the recorded measurements.

As another example, the ozone concentration level in downtown Los Angeles is measured over time to provide information about the greater 'process' defined as the air quality in Los Angeles. Inferences can be made about the effects new environmental laws have on air quality by analyzing the sequence of monthly averages of hourly ozone readings (Box and Tiao, 1975). In this case, although it is theoretically possible to continuously measure the ozone concentration level, the measurements have been taken only at discrete intervals of time. The discretization of data is a common practice in the continuous process industries (MacGregor, 1988; MacGregor and Harris, 1993) and, thus, we will assume that the performance of a process can be measured via a discrete sequence of observations taken on the process. Further, when the phrase 'the process' is used in this dissertation, it will generally refer to the measurements or observations that come from the process.

1.2.2 Process Quality. The available process measurements are assumed to relate to the quality of the process. Quality is best defined as the "fitness for use" of a product or service (Montgomery, 1991). In this context, process observations provide a measure of some quality characteristic that reflects the fitness for use of the end product. In general, the quality characteristic will have an associated target value, τ (Taguchi and Wu, 1980). This target value represents an ideal state and deviations from the target indicate lower quality. In the wing rib example, there is an ideal diameter for a drilled hole. A hole that is either too large or too small increases the risk of failure. In the air quality example, lower ozone concentrations are considered better and the ideal would be no measurable ozone, although this may be an unattainable state. Typically, the target value is finite, although two exceptions exist: smaller values may always imply higher quality or larger values may imply higher quality. In these cases, the ideal state can only be reached in a limiting sense. For simplicity, the remainder of this research will only consider the case of a finite target value.

1.2.3 Process Variation. All real processes exhibit some variation (Box and Kramer, 1992). This process variation directly implies a loss of quality due to deviations from the target value of the quality characteristic. In order to gain an understanding of the reasons for a loss of quality in a process, it is convenient to partition the total variation found in the process based upon the sources of that variation. The American Society for

Quality Control (ASQC) (1983) divides the sources of process variation into two classes: *chance cause variation* and *assignable cause variation.* Chance (or common) causes are defined by the ASQC as

> "factors, generally numerous and individually of relatively small importance, which contribute to variation, but which are not feasible to detect or identify."

On the other hand, an assignable (or special) cause is defined by the ASQC as

> "a factor which contributes to variation and which is feasible to detect and identify."

The variation in the process is manifested through the dispersion of the process observations about some value. The owner of the process may be able to discern a pattern in that dispersion, or, detect a cause of variation. By analyzing the pattern, the owner may also be able to attribute the variation to some specific source, or, identify the cause of variation. By expending enough resources (e.g., time, effort, money), the owner of the process may be able to detect and identify (almost) all of the causes of variation. Therefore, the dividing line between chance and assignable causes of variation is an economic decision made by the owner of the process and is unique to each process.

These two classes of variation provide an adequate foundation for examining simple processes. Observations from such processes exhibit only chance cause variation until some assignable cause occurs. The occurrence of an assignable cause implies the addition of assignable cause variation which is detectable in the process observations. Because assignable causes increase the variability in a process and, thus, produce a loss in quality, they typically are corrected or removed once they have been detected and identified. Indeed, although not explicitly defined as such, an assignable cause is regarded most commonly in practice as a factor that contributes to variation and which is feasible to detect, identify, *and* remove.

More complex processes exist, however, in which the observations exhibit some nonrandom structure, such as autocorrelation, which is detectable and identifiable, but which is not feasible to remove. Many such processes can be found in the continuous process industries, such as chemicals, pulp and paper, and mineral processing (MacGregor and Harris, 1990). Some authors prefer to treat the variation resulting from an autocorrelative structure as special cause variation, while other authors implicitly regard it as chance cause variation.

To avoid confusion and stay within the intent of the ASQC definitions of chance and assignable causes, a third source of variation due to structural causes is proposed in this dissertation. In the proposed taxonomy, **assignable cause variation** is the variation in the process measurements which is due to changes in the system that can be detected, identified and eliminated. For instance, a slightly bent drill bit in the wing rib example may cause an increase in the variability of the measured hole diameters. The added variation can be detected by analyzing the process measurements, traced to the drill bit, and removed by replacing the drill bit. On the other hand, **structural cause variation** is variation that can be detected and identified, but which is not feasible to remove. Structural cause variation is integral to the mechanics of the process. A good example of structural cause variation is the seasonal increase in the ozone concentration levels in the summer months. While the seasonal effects can clearly be detected and identified, it is unlikely that they can ever be removed from the process. In contrast, assignable cause variation due to loosened environmental laws can be detected, identified and removed by re-tightening the laws. In both examples, **chance cause variation** remains as the variation that is not feasible to detect or identify.

Table 1 on page 9 provides examples of a variety of processes that exhibit various combinations of the three components of variation described above. In these examples, x_t denotes the process observation made at time t and chance cause variation is included via ϵ_t, an independent observation from a normally distributed random variable with zero mean and variance equal to 0.04. Assignable cause variation is included via δ_t, a step function of magnitude 0.2 occurring after the fiftieth observation. Structural cause variation is included via f_t, a sinusoid function with magnitude $\sqrt{2}/5$ and period 100/3. In the absence of any causes of variation, the observations are a constant, μ, equal to zero. For these examples, we will use the long-term mean square deviation from zero, denoted MSD, as the measure of variation of the observations about their target. MSD is defined via

$$MSD(X) = \lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} (X_i - 0)^2. \tag{1}$$

The mean square deviation associated with the process equals 0.04 when a single source of variation is present. That is,

$$MSD(\epsilon) = MSD(\delta) = MSD(f) = 0.04. \tag{2}$$

In addition, each source is uncorrelated with the others so that, when combinations of two sources are present, the process variation increases to 0.08. When all three are present the process variation equals 0.12. That is,

$$MSD(\delta + \epsilon) = MSD(f + \epsilon) = MSD(\delta + f) = 0.08, \tag{3}$$

and

$$MSD(\delta + f + \epsilon) = 0.12. \tag{4}$$

Figure 1 on page 8 graphically depicts instantiations consisting of 100 observations for each process in Table 1 on page 9.

1.2.4 Statistical Process Control. Given that we have a measurable process which exhibits variation, Statistical Process Control (SPC) is a way of thinking and a set of tools used to improve the quality of the process (Wheeler and Chambers, 1992). In large measure, SPC aims to improve quality by reducing the variability of the process about the

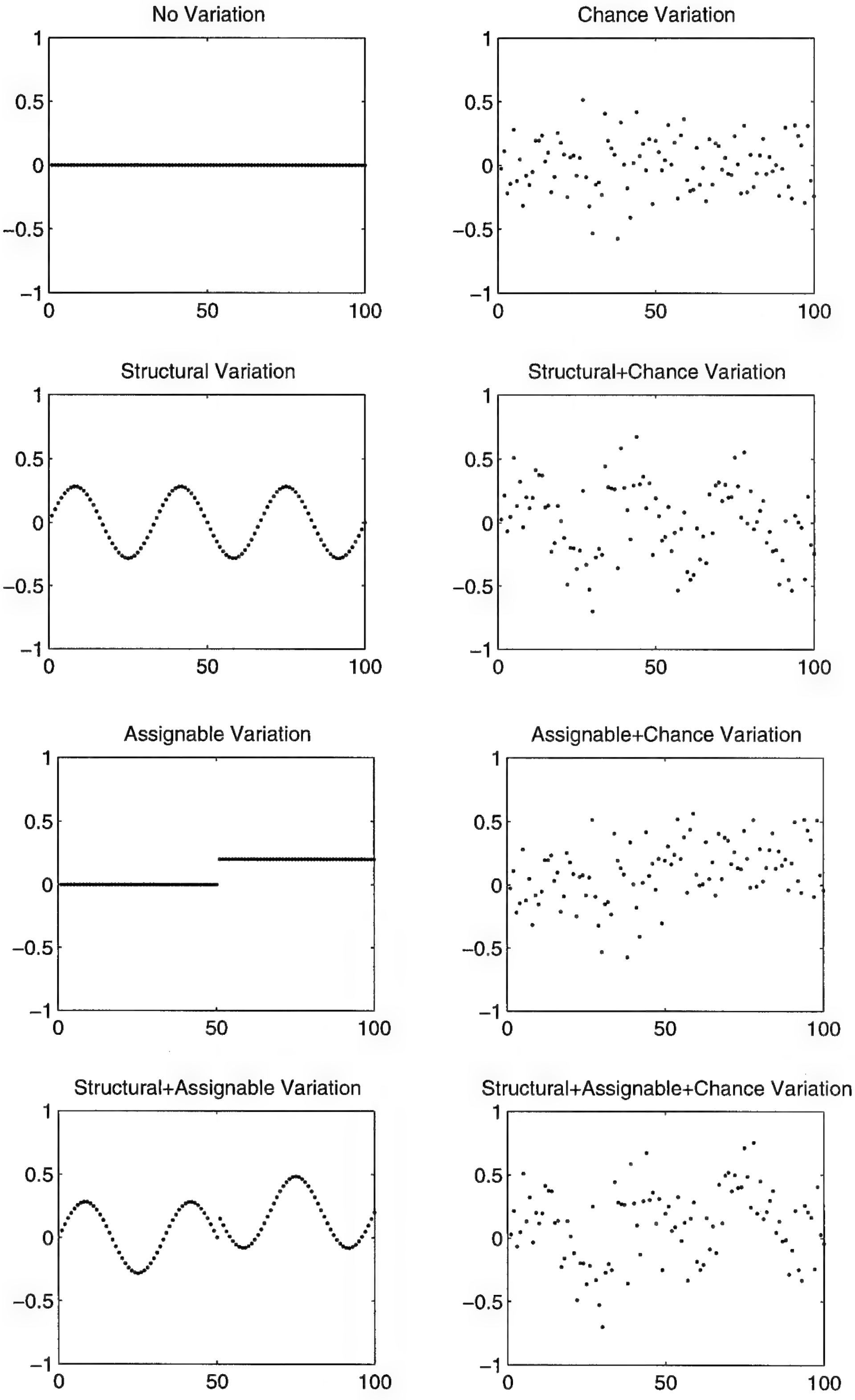


Figure 1. Instantiations of the processes in Table 1.

Table 1. Examples of processes with various causes of variation

Causes of Variation			
Chance	Assignable	Structural	Model
			$X_t = \mu$
√			$X_t = \mu + \epsilon_t$
	√		$X_t = \mu + \delta_t$
√	√		$X_t = \mu + \delta_t + \epsilon_t$
		√	$X_t = \mu + f_t$
√		√	$X_t = \mu + f_t + \epsilon_t$
	√	√	$X_t = \mu + \delta_t + f_t$
√	√	√	$X_t = \mu + \delta_t + f_t + \epsilon_t$

$$\mu = 0;$$

$$\epsilon_t \sim N(0, 1/25);$$

$$\delta_t = \begin{cases} 1/5 & \text{if } \ \mathrm{t} \geq 50 \\ 0 & \text{if } \ \mathrm{t} < 50 \end{cases};$$

$$f_t = sin(6\pi t/100)\sqrt{2}/5.$$

target value (Doty, 1990; Montgomery, 1991). Along similar lines, Beauregard, Mikulak and Olson (1992) say that SPC attempts to achieve stable, predictable process performance. The combination of a stable predictable process and reduced variability is at the heart of SPC.

In a sentence, SPC is "a statistically based approach for monitoring, controlling, evaluating, and analyzing a process" (Beauregard et al., 1992). Implicit in this definition is that action is taken to improve the process. Without action, statistical process control reduces to statistical process monitoring.

The value of SPC is found in its potential to improve the total quality produced by an organization. To understand this potential, consider statistical process control in its relationship to statistical thinking. Snee (1990) defines statistical thinking as

> "*... thought processes*, which recognize that variation is all around us and present in everything we do, all work is a series of interconnected processes, and identifying, characterizing, quantifying, controlling, and reducing variation provide opportunities for improvement."

Snee considers variation and quality as strategic concepts that must be understood in order for an organization to achieve total quality. SPC is one system of tools that align the organization's operational activities with its strategic direction.

Figure 2 on page 11, adapted from Snee, schematically depicts the application of statistical thinking to statistical process control. The diagram is intended to provide a broad understanding of the concepts and systems essential to maximizing the contribution of statistical thinking to total quality without focusing on the tools employed in implementing those systems. The diagram begins by reiterating the basic premises that 'all work is a process' and 'all processes are variable.' We include these two tenets to emphasize that SPC should be considered in terms of continuous process improvement. By analyzing the variation that must exist in the process, knowledge about the sources of the variation can be developed. Depending on the identified source of variation, one or more courses of action can be taken. In the rightmost path, the figure reflects the possibility of reducing chance cause variation by changing the process, but also implies the impossibility of totally eliminating variation. On the other hand, the leftmost path shows that assignable cause variation can be totally eliminated at a point in time by removing assignable causes. The identification and removal of assignable causes can be considered to be controlling the process, or, maintaining the process at the level of variation prior to the introduction of the assignable cause. The middle path shows that structural cause variation may be either reduced or eliminated by changing the process. Note that structural and chance cause variation cannot be reduced or removed by controlling the process. However, process knowledge gained by applying statistical thinking may enable the owner of the process to change to a new process with less structural or chance cause variation. All three paths lead to a reduction in process variation and, hence, an improvement in process quality. Finally, the figure portrays the iteration required to continually improve the process.

1.2.5 The State of Statistical Control. Walter Shewhart is credited with originating the field of Statistical Process Control. He was primarily interested in maintaining a process in a state of control and said that

> "a phenomenon will be said to be controlled when, through the use of past experience, we can predict, at least within limits, how the phenomenon may be expected to vary in the future. Here it is understood that prediction within limits means that we can state, at least approximately, the probability that the observed phenomenon will fall within the given limits."

At the time Shewhart first published his ideas in 1939 this was a grand and novel concept. While mass production techniques had existed for over a century, the standardization of production was just entering the picture (Shewhart, 1986). The standards imposed upon the production processes demanded some method for "minimizing the number of rejections" while at the same time "minimizing the cost of inspection required to give adequate assurance of quality." Shewhart saw the need to bring the production systems of his day into a state of control as a first step toward improving their quality and profitability.

More recently, the American Society for Quality Control (1983) defined a process to be in a state of statistical control "if the variations among the observed sampling results (from the process) can be attributed to a constant system of chance causes." Chance causes refer to the built-in chronic variation found in the process. It will remain as a part of the process unless the process is changed. By a constant system, the definition implies that the chance causes occur in a manner that does not vary with time. The strictest and most widely accepted interpretation of the ASQC definition is that the observations from an in-control process should be independent and identically distributed (iid) with a constant mean and variance.

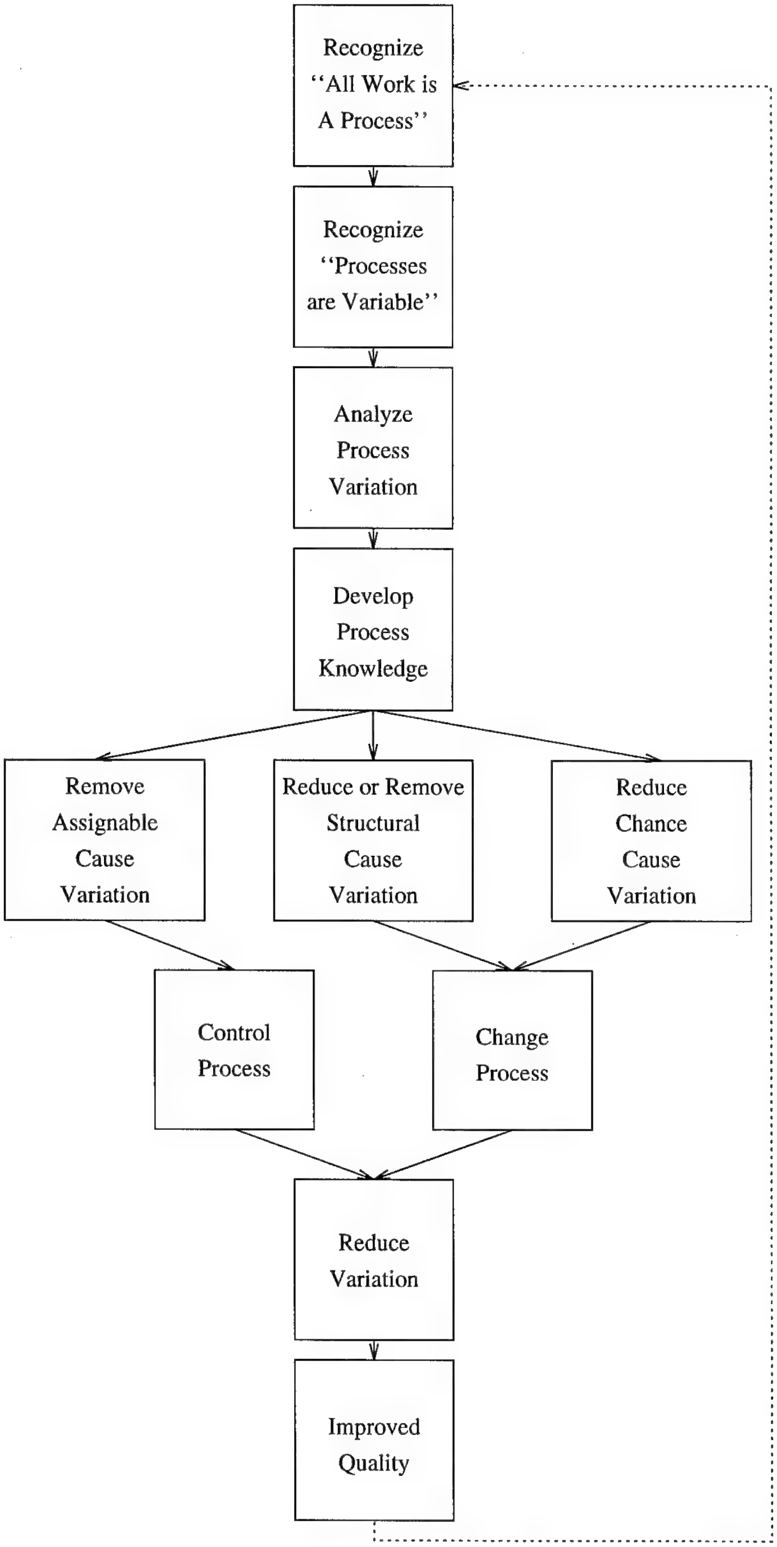

Recognize "All Work is A Process"
Recognize "Processes are Variable"
Analyze Process Variation
Develop Process Knowledge
Remove Assignable Cause Variation
Reduce or Remove Structural Cause Variation
Reduce Chance Cause Variation
Control Process
Change Process
Reduce Variation
Improved Quality

Figure 2. Statistical thinking in quality improvement

A corollary definition also exists for the state of being out of statistical control. A process is said to be out of statistical control when there exists some variation in the process that can not be attributed to a constant system of chance causes. That is, some variation in an 'out-of-control' process can be attributed to either an assignable or structural cause of variation. More recent ideas that challenge and expand these interpretations are discussed in the next chapter.

Figure 3 illustrates the counterintuitive nature of the definition of control. It shows 100 observations from two simulated processes. In one of the processes, the observations arise from a Normal distribution with a mean of 0 and a variance of 1, which is sometimes referred to as a white noise process. The other process consists of discrete samples from a simple sinusoid centered at 0 with an amplitude of $1/\sqrt{2}$ and a period of 2π. Under the strict interpretation of the ASQC definition, it is clear that the white noise process is 'in control' since all of the variation exhibited by the process is due to independent and identically distributed errors. The sinusoid, however, exhibits structural cause variation that is not attributable to a constant system of chance chances. The sinusoidal process is therefore 'out of control' according to a strict interpretation of the ASQC definition. On the other hand, future values of the sinusoid process can be quite accurately predicted and so, by Shewhart's original definition, the sinusoid process is also 'in control.' These two cases illustrate the care that must be taken when using the phrase 'in control.' For the remainder of this dissertation, the ASQC definition of control will be used.

The use of one of the key techniques used in Statistical Process Control, the control chart, implicitly relies on the interpretation that the observations from an 'in-control' process are independent and identically distributed. In this case, the addition of an assignable cause of variation to the process should be reflected in observations that do not fit the in-control distribution. That is, the change should be reflected in observations that do not appear to be samples from the in-control distribution or do not appear to be independent. Using statistical techniques, the observed changes in process variation can be quantified

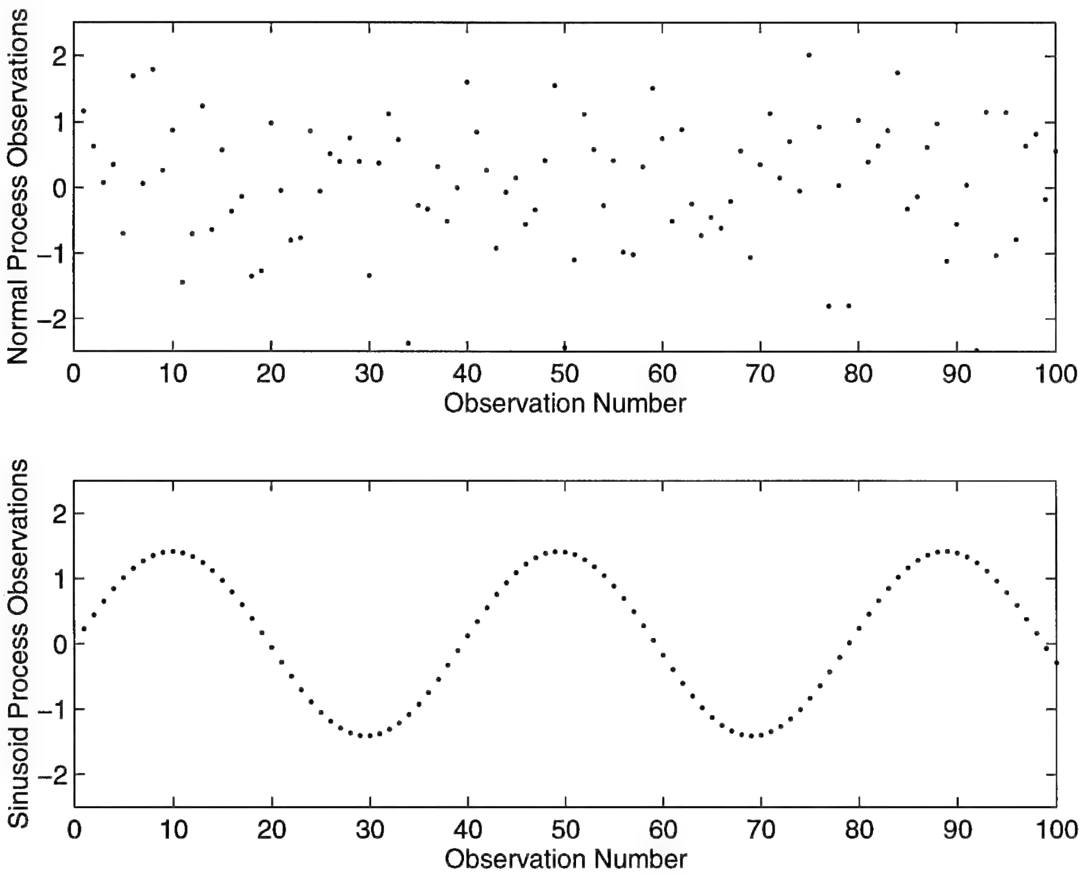


Figure 3. Samples from a Normal(0,1) Distribution and a Sinusoid Function

and an assessment can be made as to whether or not the process is still 'in control.' If there is sufficient statistical evidence that process observations no longer match the in-control distribution, we conclude that the process is 'out of control.' The graphical display of this type of statistical test for the state of control is commonly referred to as a control chart.

1.2.6 Capability. The control chart aims to improve quality by maintaining a process in a state of statistical control and thus controlling the variation exhibited by the process. While the control chart indicates the state of statistical control, it does not indicate how suitable the output of the process is for its intended purpose. Process capability is a measure of suitability, and hence, also measures the quality of a process. In addition to the previously discussed target value, a process will generally have associated Upper and Lower Specification Limits (USL and LSL) which define the range of quality characteristic values that are acceptable for use by the owner of the process. Any output from the process whose quality characteristic lies outside of the specification range is deemed to be unacceptable for use. In a manufacturing environment, that condition results in scrap or rework.

Beauregard defines process capability as a "measure of the total variation in the process output against the specifications." All else being equal, a process with a higher variability will produce more items outside of its specification limits, and so will be said to be less capable. A capable process generates (almost) all of its output inside of its specification range. As a rule of thumb, a capable process produces not more than 0.1% outside of its specification limits (Bissell, 1990).

Wheeler and Chambers (1992) connect the concepts of control and capability. When a process is operating in a state of statistical control, it is possible to quantify the variation that exists in that process and, therefore, to quantify the capability of the process. They point out that any stable process can be said to possess a well-defined capability, although the variation in the process may be quite large compared to the specifications. Further, they assert that a reliable prediction of future variation cannot be made for a process which is currently subject to assignable causes of variation. This unpredictability leads to their conclusion that a process must be reasonably in-control before it can be considered capable.

1.3 Chapter Summary

In this chapter, we introduced basic concepts that are used throughout this research. These concepts are unified by the objective to maximize process quality. Statistical thinking provides the framework to achieve that objective. Statistical thinking recognizes that all processes exhibit some variation and this variation directly leads to a reduction in the quality derived from the process. An improved understanding of the process can be gained by analyzing the variation exhibited by the process. Understanding the causes of variation allows those causes to by controlled or eliminated and, therefore, allows quality to be improved.

We propose a taxonomy of three causes of variation to aide in understanding complex processes. The three causes of variation are chance causes, assignable causes and structural causes. By analyzing the statistical properties of a process in terms of these causes of variation and taking the appropriate corrective actions, process variation can be reduced and product quality improved.

Statistical process control is one set of tools used to improve quality. An important assumption used in many SPC techniques is that quality improvement can be achieved by maintaining the process in a state of statistical control. In Chapter II, we review currently accepted SPC techniques and discuss their limitations in the presence of autocorrelated observations. In Chapter III, we develop a new technique for selecting control limits for low order models. Finally, we develop new techniques for quality improvement based upon monitoring process capability, rather than the state of statistical control, in Chapters IV and V.

II. Background

In Chapter I, we introduced the conceptual framework for improving quality with SPC. In this chapter, we provide a detailed review of generally accepted SPC techniques which are in use today. The first step in our review is an examination of the design and assumptions behind the standard control charts. We highlight the limitations of the classic techniques when applied to autocorrelated observations and present some techniques proposed in the literature for addressing those limitations. We also discuss current techniques for quantifying the capability of a process. Finally, we give some background on the economic aspects of SPC.

2.1 Standard Control Charts

The basic SPC tool is the control chart. The standard control charts that we discuss in this section are designed to sequentially test the hypothesis that a process is in a state of statistical control versus the alternative that it is not. We generally assumed that when a process is in a state of statistical control, the observations from that process will appear as independent and identically distributed samples from a distribution with a fixed mean, μ_0, and a fixed positive standard deviation, σ_0. If the true underlying distribution is known, then an exact statistical test can be developed which enables us to infer whether or not a given sample is from the known distribution. The graphical display of this statistical test over time is called a control chart.

2.1.1 The Design of Tests for SPC. The null hypothesis for a standard control chart is that the process is in a state of statistical control. According to the definition of statistical control, the variations among the observations taken from an in-control process are attributable to a constant system of chance causes. Even though Shewhart's definition of control is less restrictive than the ASQC definition, the tools he developed rely upon the assumptions found in the strict interpretation of the ASQC definition. If we let x_t denote the observation recorded at time t from a random variable X_t, then the standard assumptions made in developing the statistical tests that underlay standard control charts can be given as

$$\begin{aligned} E(X_t) &= \mu \\ Var(X_t) &= \sigma^2 \\ Cov(X_t, X_{t+k}) &= 0 \;\; \forall \, k \neq 0 \end{aligned} \tag{5}$$

where μ and σ^2 are the mean and variance of the process, respectively. Most SPC control charts are designed to signal an out of control event when the one of the first two conditions in expression 5 is deemed to be statistically unlikely given the evidence provided by sampled observations. The control chart does not test the third condition; it is implicitly assumed to hold.

The design of a particular statistical test generally requires a tradeoff between two risks. First, when the process is truly in control, we would like the probability of falsely concluding that the process is out of control to be as low as possible. Rejecting a null hypothesis when it is true is referred to as a type I error. We generally denote the probability of a type I error as α and refer to α as the level of significance. In contrast, the power of a statistical test is a measure of how well the test will correctly determine that the null hypothesis is false. For instance, if the variance exhibited by the process increases due to the occurrence of an assignable cause, then the process is out of control and we would like the probability of detecting the change to be as high as possible. A type II error is committed if we "accept" (i.e., fail to reject) the null hypothesis that the process is in control when, in fact, it is not. The probability of a type II error is denoted as β, while the power of the test is defined to be $(1 - \beta)$. Adjustments to test parameters, such as the sample size, which increase the test's power will also tend to increase the test's level of significance, and vice versa.

In general, a fixed level of significance is specified and the test parameters are chosen to maximize the power of the test (Montgomery, 1991).

In the simplified environment defined by expression 5, the probability of a type II error is closely tied to one of two types of changes that may take place in the process. The first is a change in the process mean, also called a mean shift. The second is a change in the process standard deviation.

A standard method for reporting how well a particular test performs is to tabulate the average run length (ARL) of the test for a variety of conditions. The average run length is defined as the expected number of samples tested before a shift in the process is signaled (Aroian and Levene, 1950; American Society for Quality Control, Statistics Division, 1983; Page, 1954). Notationally, when a process parameter (say the mean) has shifted by an amount, Δ, the average run length immediately after the change is given as ARL(Δ). When the process is in control, its average run length is given as ARL(0). Naturally, we would like the average run length when the process is in control to be large; we do not want to falsely signal an out of control situation very often. Similarly, we would like the average run length to be small when the process is out of control. These two preferences are analogous to minimizing the level of significance while maximizing the power of the test, respectively.

2.1.2 ` Estimating μ_0 and σ_0. When the exact underlying distribution of the in-control observations is known, well established statistical tests can be used to construct tests for the state of control. Unfortunately, the true underlying distribution is generally not known. A first step toward constructing the standard control charts is to estimate the mean and the standard deviation of the true distribution from a set of observations which are assumed to be in control.

Suppose we have m subgroups of process observations, each of equal size, n. The mean of each subgroup, $\bar{x}_i$, can be computed via

$$\bar{x}_i = \frac{1}{n}\sum_{j=1}^{n} x_{ij}. \tag{6}$$

The mean of the underlying distribution, μ_0, can then be estimated with the grand mean,

$$\bar{\bar{x}} = \frac{1}{m}\sum_{i=1}^{m} \bar{x}_i. \tag{7}$$

The standard deviation is usually estimated in one of two ways. The first estimate of the standard deviation is the average standard deviation of the subgroups. That is, the standard deviation of each sample is first calculated via

$$s_i = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_{ij} - \bar{x}_i)^2}. \tag{8}$$

Then an estimate of σ_0 is given by the mean of the subgroup standard deviations, $\bar{s}$,

$$\bar{s} = \frac{1}{m}\sum_{i=1}^{m} s_i. \tag{9}$$

It is a fact that the grand mean is an unbiased estimator of the true mean (DeGroot, 1989, pg 412):

$$E(\bar{\bar{x}}) = \mu_0. \tag{10}$$

It is also a fact that s^2 is an unbiased estimator of σ_0^2 (DeGroot, 1989, pg 413). However, Montgomery (1991) states that s is a biased estimator of σ_0. Further, when the underlying distribution is normal,

$$E(s_i) = \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]}\sqrt{\frac{2}{n-1}}\,\sigma_0 \tag{11}$$

where $\Gamma(\cdot)$ is the "complete gamma function" defined by

$$\Gamma(a) \equiv \int_0^\infty x^{a-1} e^{-x} dx. \tag{12}$$

A constant, c_4, which is a function of the sample size, n, is defined as

$$c_4 \equiv \frac{\Gamma(n/2)}{\Gamma[(n-1)/2]} \sqrt{\frac{2}{n-1}}. \tag{13}$$

Thus, $\bar{s}/c_4$ is an unbiased estimator of σ_0, i.e.,

$$E(\bar{s}/c_4) = \sigma_0. \tag{14}$$

The average range provides an alternative method for estimating σ_0. The range of a subgroup is defined as

$$R_i = \max_{j=1,\dots,n} \{x_{ij}\} - \min_{j=1,\dots,n} \{x_{ij}\} \tag{15}$$

and the average range is defined as

$$\bar{R} = \frac{1}{m} \sum_{i=1}^{m} R_i. \tag{16}$$

Like $\bar{s}$, the average range is biased. Fortunately, when the underlying distribution is normal, there exists a constant, d_2, corresponding to c_4 such that $\bar{R}/d_2$ is an unbiased estimator of σ_0, i.e.,

$$E(\bar{R}/d_2) = \sigma_0. \tag{17}$$

Tabled values of c_4 and d_2 for various sample sizes can be found in most SPC textbooks (e.g. Appendix VI of Montgomery).

2.1.3 $\bar{X}$-Chart. Estimates of the mean and standard deviation for the in-control process are used to construct control charts. The most common control chart is the $\bar{X}$-

chart. Ideally, we would like to test the hypothesis that the process is in control and, hence, that the observations arise from independent and identically distributed samples. In an $\bar{X}$-chart, however, we only test the more restrictive null hypothesis that the mean of the observations is equal to the mean of the in-control distribution, μ_0:

$$H_0 : \mu = \mu_0$$

$$\text{versus}$$

$$H_A : \mu \neq \mu_0. \tag{18}$$

Suppose the process is in control. Then, under the strict interpretation of the ASQC definition of statistical control, the observations will be independent and identically distributed. In addition, when the observations are normally distributed, the means of the subgroups will also be normally distributed. Although the theory underlying the $\bar{X}$-chart is based on normal theory, normally distributed averages are not a prerequisite of their use. According to the Central Limit Theorem, the sample averages will have a limiting distribution that is normal (Hogg and Craig, 1978, pg 193). Schilling and Nelson (1976) also report that, for most practical applications, samples of size four or more ensure that sample averages follow the approximately normal distribution required of an $\bar{X}$-chart. Although the sample averages are approximately normally distributed, the parameters of the approximate distribution are unknown. A natural estimate for μ_0 is the grand mean, $\bar{\bar{x}}$. Similarly, σ_0 can be estimated with either $\bar{s}/c_4$ or $\bar{R}/d_2$. The hypothesis given above requires a two-sided test. An approximate critical region for this test is the area outside of the range:

$$\mu_0 \pm k_{\alpha/2} \, \frac{\sigma_0}{\sqrt{n}} \tag{19}$$

or, with the estimates substituted in:

$$\bar{\bar{x}} \pm k_{\alpha/2} \, \frac{\bar{s}}{c_4\sqrt{n}} \tag{20}$$

where α is the desired level of significance for the test and $k_{\alpha/2}$ is a corresponding value from the cumulative density function of the normal distribution, denoted $\Phi(\cdot)$, such that:

$$\Phi(k_{\alpha/2}) = 1 - \frac{\alpha}{2}. \tag{21}$$

When the average of a subgroup falls outside of this range, there is statistical evidence that the observations do not come from a distribution with a mean of μ_0. We refer to the endpoints of the critical region as the upper and lower control limits (UCL and LCL). The plot of this test over time it is referred to as an $\bar{X}$ chart. An example of an $\bar{X}$ chart is depicted in Figure 4. In practice, a value of 3 is frequently specified for $k_{\alpha/2}$ and the resulting control limits are referred to as 3σ control limits. Since the test statistic defined above relies upon a constant standard deviation, the $\bar{X}$ chart is also somewhat sensitive to changes in the standard deviation. Additionally, the assumption of normality due to the Central Limit Theorem is relied upon in practice while, in theory, a more correct test based upon the t-distribution could be called for, especially when estimates of parameters are used.

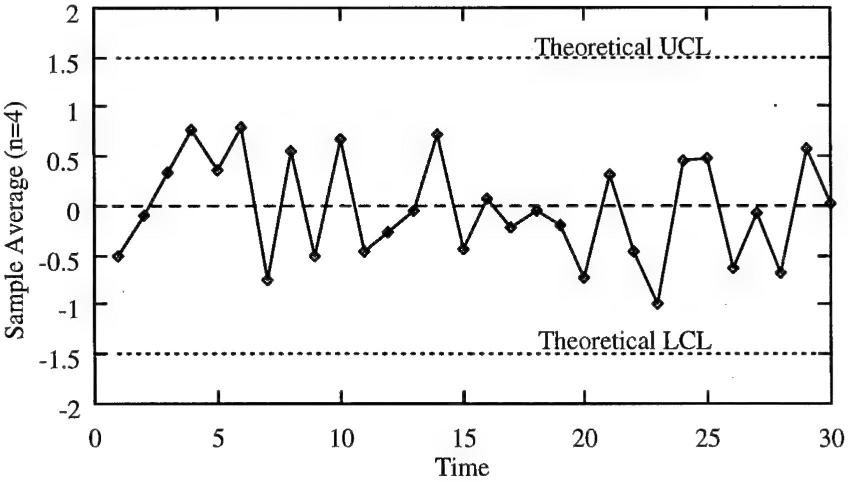


Figure 4. $\bar{X}$-chart for white noise with samples of size 4

2.1.4 S and R Charts. The $\bar{X}$-chart explicitly tests the hypothesis that the mean of the sample is equal to the in-control mean. Since the normal distribution is defined by two parameters, the mean and the variance, we can also test the hypothesis that the variance or standard deviation is constant, that is:

$$H_0 : \sigma = \sigma_0$$

versus

$$H_A : \sigma \neq \sigma_0. \tag{22}$$

We have already seen two methods for estimating σ_0 for the $\bar{X}$-chart. Each method has a corresponding control chart. Recall that, when the observations are normally distributed,

$$E(s) = c_4 \sigma_0. \tag{23}$$

Montgomery (1991) further states that

$$Var(s) = (1 - c_4^2)\ \sigma_0^2. \tag{24}$$

Then control limits may be constructed which have the form:

$$c_4 \sigma_0 \pm k_{\alpha/2}\ \sigma_0 \sqrt{1 - c_4^2} \tag{25}$$

or, with the estimator substituted:

$$\bar{s} \pm k_{\alpha/2}\ \bar{s} \frac{\sqrt{1 - c_4^2}}{c_4}. \tag{26}$$

This control region is the basis of the S chart. The R chart is very similar. Recall that, when the observations are normally distributed,

$$E(\bar{R}) = d_2\ \sigma_0. \tag{27}$$

In addition, the variance of $\bar{R}$ can be expressed in terms of d_3, a known function of the sample size n, via

$$Var(\bar{R}) = d_3^2\, \sigma_0^2. \tag{28}$$

The control region for the R chart is then determined to be

$$\bar{R} \pm k_{\alpha/2}\, d_3 \frac{\bar{R}}{d_2}. \tag{29}$$

Tabulated values of d_2 and d_3 for various sample sizes can be found in most SPC textbooks (e.g. Appendix VI of Montgomery).

2.1.5 Individuals Chart. It may be undesirable or infeasible to group the process observations. In this case, the $\bar{X}$-chart is no longer appropriate for monitoring control. In particular, the two previously described estimates of the process standard deviation ($\bar{s}/c_4$ and $\bar{R}/d_2$) require subgroups with more than one observation. Additionally, when only one observation is available per group, the central limit theorem can no longer be applied to ensure the approximate normality of the sample averages. The individuals chart therefore assumes that the distribution of the individual observations themselves is, at least approximately, normal. In order to create a control chart for individuals, called an X-chart, we can estimate σ_0 as the standard deviation of the entire set of data from the in control process with s via

$$s = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n}(x_j - \bar{x})^2} \tag{30}$$

where n is the total number of observations. The moving range is an alternative estimator for σ_0. The moving range is defined as

$$MR_i = |x_i - x_{i-1}| \tag{31}$$

and the average moving range is defined as

$$\overline{MR} = \frac{1}{n-1}\Sigma_{i=2}^{n} MR_i \tag{32}$$

where n is the total number of available observations. It should be noted that $\overline{MR}$ possesses a bias similar to that of $\bar{R}$. The same bias correction applied to $\bar{R}$ can be applied to $\overline{MR}$ by dividing $\overline{MR}$ by d_2. Similarly, the simple mean is used as an estimate of μ_0:

$$\bar{X} = \frac{1}{n}\Sigma_{i=1}^{n} x_i. \tag{33}$$

Using $\bar{X}$ as an estimate of μ and $\overline{MR}/d_2$ as an estimate of σ, a control region for the X-chart is defined as

$$\bar{X} \pm k_{\alpha/2}\frac{\overline{MR}}{d_2}. \tag{34}$$

2.1.6 MR Chart. Just as the $\bar{X}$ chart had a companion chart to test for changes in the variance (the S or R charts), the X chart has a companion chart (the moving range or MR chart). The centerline for the MR chart is naturally chosen to be the average moving range, $\overline{MR}$. The distribution of $\overline{MR}$ is not simple. Fortunately, factors D_3 and D_4 are available (see Montgomery, Appendix VI) to construct a control region for the MR chart via

$$D_3\,\overline{MR} \leq MR_j \leq D_4\,\overline{MR} \tag{35}$$

where MR_j is the observed moving range that is being tested. The factors D_3 and D_4 provide a control region that is equivalent in power to a 3σ control region for an $\bar{X}$-chart, when the observations are independent and normally distributed.

2.1.7 Supplementary Runs Rules. One method for improving the power of a standard control chart is to augment the chart with one or more supplementary runs rules. Recall from Section 2.1.3 that a standard rule for the $\bar{X}$-chart is to generate an out of control signal whenever a sample mean falls outside of the 3σ control limits. Graphically, the expected pattern of sample means clustering around the grand mean is broken by a sample mean that is too far away from the grand mean. When a process is in control and its observations are independent and identically distributed, no pattern other than a general clustering and fluctuation around the centerline is expected in the control chart. A supplementary runs rule is simply an additional rule for signaling an out of control condition based upon detecting patterns that are unlikely to occur when the plotted points are independent and identically distributed.

Champ and Woodall (1987) discuss the most generally accepted supplementary runs rules. These include signaling an out-of-control condition when one or more of the following occur:

- eight plotted points in a row are to one side of the center line,
- two plotted points out of three are between either 2σ and 3σ or -2σ and -3σ,
- four plotted points out of five are between either 1σ and 3σ or -1σ and -3σ.

Note that these are just some possible supplementary runs rule and that the rules are not themselves independent. These particular rules were developed in the context of enhancing the $\bar{X}$-chart. Also note that the addition of a supplementary runs rule will increase the probability of a false alarm and, therefore, decrease the average run length. For example, applying the three listed rules individually to a normally distributed in-control process with 3σ control limits causes a decrease in the average run length from 370.4 to 152.7, 225.4 and 166.1, respectively. Adding all three rules to the original process decreases its average run length to 91.8.

2.1.8 CUSUM Chart. One of the problems with the $\bar{X}$ chart is that it is insensitive to small shifts in the mean. When the mean shift is less than $\pm\sigma_0$, the average run length does not drop significantly (Montgomery, 1991). The cumulative sum (CUSUM) control chart is designed to detect a shift in the mean but is based upon a different statistical test.

In practice, the CUSUM chart performs better than the $\bar{X}$ chart for detecting small shifts in the mean.

The CUSUM test was first described by Page (1954, 1955). The basic idea behind the CUSUM test is that the cumulative effects of a small shift in the process mean can be detected more quickly than waiting for a single sample average to be extreme enough to generate a signal from the $\bar{X}$ chart.

A two sided CUSUM is used when the parameter may shift in either direction. The CUSUM statistics at time t are defined as

$$\begin{aligned} S_H(t) &= \max[0, \bar{x}_t - (\mu_0 + K) + S_H(t-1)] \\ S_L(t) &= \max[0, -\bar{x}_t + (\mu_0 - K) + S_H(t-1)] \end{aligned} \tag{36}$$

with the starting values $S_H(0) = S_L(0) = 0$. $\bar{x}_t$ is the average of the subgroup sample taken at time t. K is called the reference value and is usually chosen to be about one-half of the size of shift to be detected. An out-of-control condition is signalled when either of the CUSUM statistics exceeds a critical value which is based on the size of the shift to be detected.

Lucas and Crosier (1982a) propose a modification to the CUSUM to permit a more rapid response to an initial out of control situation. In their scheme, $S_H(0)$ and $S_L(0)$ are set to some initial 'head-start' value. When the process is in control, the CUSUM statistics will tend to return to 0. However, when the process is out of control, the modified CUSUM statistic will tend to exceed the reference value more quickly than an unmodified CUSUM. In a later paper (1982b), they discuss robust procedures for handling outliers with the CUSUM.

2.1.9 EWMA Charts. Like the CUSUM chart, the exponentially weighted moving average (EWMA) control chart was designed to detect smaller shifts in the mean more quickly than the Individuals and $\bar{X}$ charts. The EWMA statistic is defined as

$$z_t = \lambda \bar{x}_t + (1 - \lambda) z_{t-1} \tag{37}$$

where $0 < \lambda \leq 1$ is a constant. The starting value, z_0, is usually chosen to be the mean of the observations, $\bar{\bar{x}}$. Montgomery (1991) shows that the control limits for the EWMA are

$$\begin{aligned} UCL &= \bar{\bar{x}} + A_2 \bar{R} \sqrt{\frac{\lambda}{2 - \lambda}} \\ LCL &= \bar{\bar{x}} - A_2 \bar{R} \sqrt{\frac{\lambda}{2 - \lambda}} \end{aligned} \tag{38}$$

where A_2 (see Montgomery, Appendix VI) is a factor to convert subsample ranges to 3σ ranges, when the observations are independent and normally distributed.

Roberts first proposed the EWMA control scheme in 1959. More recently, Lucas and Sacucci (1990) provided design considerations for parameter selection. They also discuss enhancements to the EWMA control scheme including a combined Shewhart-EWMA, a robust EWMA, and a fast initial response feature. They conclude that average run length characteristics of the EWMA are comparable to the CUSUM control scheme.

2.1.10 Use of Control Charts in Practice. The most widely used control charts are the $\bar{X}$ and R charts. According to Montgomery (1991), "the $\bar{X}$ and R (or S) charts are among the most important and useful on-line statistical process control techniques." All of the control charts discussed so far are appropriate when process observations are, at least approximately, independent and normally distributed. Furthermore, the control charts can be easily implemented using only paper, pencil and calculator. The EWMA and CUSUM charts are generally only used when it is important to detect relatively small shifts in the mean (i.e. shifts of less than one standard deviation.)

2.2 Relaxing the Assumption of Independence

The standard control charts we described in the previous section all implicitly assume that the process observations are independent. However, Box and Kramer (1992) make the assertion that, in their experience, "all processes are autocorrelated." Certainly, there are a large number of case studies in which real world data sets exhibit autocorrelation. This naturally raises the question of how well the standard control chart techniques perform with autocorrelated processes. We describe some attempts to address that question in this section.

Over the past decade, a significant portion of the research documented in the statistical quality control literature has challenged the assumption of independence. In most of those studies, the independence assumption is replaced by an assumption that the process can be adequately modeled by a particular type of autocorrelated time-series model. The effects of the autocorrelation specified in the time-series model account for the dependence found in the original observations.

2.2.1 The Effects of Autocorrelation. The correlation between two random variables X and Y with finite variances σ_X^2 and σ_Y^2, is denoted by $\rho(X,Y)$ and defined as (Hogg and Craig, 1978)

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}. \tag{39}$$

Correlation is a measure of the tendency of two random variables to vary linearly together. For a stationary time-series, X_t, with mean μ_X and finite variance σ_X^2, the autocorrelation at lag k is defined as (Box and Jenkins, 1976):

$$\begin{aligned} \rho_k &= \frac{E[(X_t - \mu_x)(X_{t+k} - \mu_x)]}{\sqrt{E[(X_t - \mu_x)^2]E[(X_{t+k} - \mu_x)^2]}} \\ &= \frac{E[(X_t - \mu_x)(X_{t+k} - \mu_x)]}{\sigma_X^2}. \end{aligned} \tag{40}$$

Autocorrelation is a measure of the tendency of neighboring observations from a time-series to vary linearly together rather than independently.

Maragah and Woodall (1992) show that, in the presence of positive first-lag autocorrelation, $\bar{R}/d_2$ provides a biased estimate for σ_0. They derive the equality

$$E(\bar{R}/d_2) = \sigma_x \sqrt{1-\rho_1} \tag{41}$$

for a stationary process where σ_x is the process variation and ρ_1 is the first lag autocorrelation. Clearly,

$$E(\bar{R}/d_2) < \sigma_x \quad \text{for} \quad 0 < \rho_1 < 1 \tag{42}$$

and

$$E(\bar{R}/d_2) > \sigma_x \quad \text{for} \quad -1 < \rho_1 < 0. \tag{43}$$

The effect of using $\bar{R}/d_2$ to estimate σ_x when $\rho_1 > 0$ is to narrow the control region. This, in turn, will result in a higher than expected false alarm rate. The opposite effect will occur for negative first lag correlation.

2.2.2 Engineering Process Control Approaches. Engineering process control is a significant alternative strategy to statistical process control for quality improvement. Engineering process control specifically addresses processes which exhibit autocorrelation. Statistical process control approaches tend to intervene in the process only when some statistical evidence indicates a source of removable or reducible variation. Engineering process control approaches, on the other hand, adjust the process after every observation. Rather than attempting to remove sources of variation, engineering process control approaches attempt to reduce process variation "by transferring the variability in the output to an input control variable" (Montgomery et al., 1994). The operation of a household thermostat in winter provides a good example of an engineering process control system. Variation from the ideal temperature of 68 degrees Fahrenheit is reduced by turning on and off the furnace. A statistical process control approach might recognize a persistent drop in temperature has occurred, prompting for a search for an assignable cause, and result in the closing of an open window. Some recent work has proposed a unified approach combining engineering and statistical process control (Montgomery et al., 1994; Box and Kramer, 1992; Vander Wiel et al., 1992; MacGregor, 1988).

Engineering process control approaches are not without their problems. A significant amount of human expertise may need to be required to design and implement the control system. It may also be too expensive or infeasible to adjust the process after each observation. In addition, frequent adjustments may mask the influence of a significant assignable cause of variation.

2.2.3 Model Fitting Approaches. Another proposed method for dealing with correlated data is fitting an appropriate time-series model to the data. The residuals from an adequately fit model may be treated as independent and identically distributed observations. The standard control chart approaches can then be applied to the residuals. Alwan and Roberts (1988) describe this concept. They call the control chart of the residuals a Special-Cause Chart (SCC) and the control chart of the fitted values a Common-Cause Chart (CCC). Each point in the Common-Cause Chart is an estimate of the local level of the process. Alwan and Roberts propose that the Common-Cause Chart can be used to help determine when a process ought to be re-centered. Montgomery and Friedman (1989) report that applying standard control charts to the sequence of residuals is effective at detecting shifts in both the location and dispersion in the original process.

The autoregressive integrated moving average (ARIMA) class of models is frequently chosen as a time-series model because it has been shown to be capable of modeling, at least approximately, the behavior of a large variety of processes. The ARIMA(p,d,q) model has the form

$$\Phi_p(B)\nabla^d X_t = \Theta_q(B)\epsilon_t \tag{44}$$

where B is the backshift operator $(B\,X_t = X_{t-1})$, $\Phi_p(B) = (1-\phi_1 B - \phi_2 B^2 - \ldots - \phi_p B^p)$ is the autoregressive polynomial of order p, $\Theta_q(B) = (1-\theta_1 B - \theta_2 B^2 - \ldots - \theta_q B^q)$ is the moving average polynomial of order q, ∇ is the backward difference operator $(\nabla^d = (1-B)^d)$, and ϵ_t is a sequence of normally and independently distributed random "shocks" with mean 0 and constant variance σ_ϵ^2. Box and Jenkins (1976) show that the residuals from an appropriately identified and fit ARIMA model will behave like independent and identically distributed random variables. Some useful examples of the ARIMA model include the ARIMA(p,0,0) or AR(p) model; the ARIMA(0,0,q) or MA(q) model; the ARIMA(p,0,q) or ARMA(p,q) model; and the ARIMA(0,1,1) or IMA(1,1) model.

Several studies have demonstrated the usefulness of the ARIMA model fitting approach for process control. Bagshaw and Johnson (1977) use a CUSUM chart on the residuals from an IMA(1,1) model to detect changes in the underlying process governing IBM stock prices. Berthouex, Hunter and Pallsen (1978) fit a seasonal ARIMA model to Sewage Treatment Plant data and successfully use the residuals in a control chart. Ermer, Chow and Wu (1979) fit the data from a nuclear reactor using an ARMA(n, n-1) model. They show that, for their highly autocorrelated data, a chart based on the sum of square residuals is much more sensitive to changes in the process than a standard control chart. Yourstone and Montgomery (1989) similarly propose a test based upon examining the residuals from a moving window of 50 observations using a known initial low order ARMA model. The residuals from the known model should be independent and normally distributed and, therefore, should not exhibit significant autocorrelation. Yourstone and Montgomery argue that a change in the process will be reflected by significant sample autocorrelation and, thus, can be tested for. They further elaborate that their proposed control chart is more sensitive to small changes in the process mean than equivalent standard control charts.

Model fitting approaches also have associated problems. Fitting models is expensive, especially in time and required expertise. In addition, since the effects of a significant assignable cause of variation may only be reflected in a single residual, the assignable cause may never be detected if that residual does not cause a signal from the control chart. The assignable cause is most likely to be detected during a window of opportunity and, after that window, the effects of the assignable cause may be incorporated by the model. Other assignable causes may be reflected by a series of relatively small changes in the residuals that are individually unlikely to cause a signal from the control chart.

2.2.4 EWMA Approaches. An alternative to fitting a model is the use of an EWMA control chart. This approach is proposed by Montgomery and Mastrangelo (1991). They point out that the EWMA is actually a subset of the ARIMA model fitting approach. The EWMA (with a properly selected parameter) provides an optimal one-step ahead forecast for the IMA(1,1) model. They cite other research that concludes the AR(1) model is also well-predicted by the EWMA.

More recently, Wardell, Moskowitz and Plant (1992) compared the EWMA control chart with the Special-Cause Chart and Shewhart control chart. They examined how quickly each chart would detect a mean shift in an ARMA(1,1) model as measured by the average run length. By varying the parameters of the ARMA(1,1) model, they tested many interesting sub-models, including the AR(1), MA(1) and Normal(0,1) models. They concluded that for a large range of models the EWMA provided better detection capabilities than either the Special-Cause Chart or the Shewhart control chart.

Choosing the parameters for an EWMA control chart requires detailed knowledge about the model. It this regard, the EWMA approach incurs the same expenses associated with model fitting approaches. In addition, the EWMA is specifically designed to detect shifts in the process mean, not in the process variance. It may not be a good technique when the impact of an assignable cause is unknown.

2.2.5 Recap. The presence of autocorrelation in process observations has been shown to cause problems with standard control charts. For example, an X-chart constructed with $\pm 3\sigma_\epsilon$ control limits on an autocorrelated process will have a higher false alarm rate, and hence, shorter average run length, than would be expected for a process with independent and identically distributed observations. In Chapter III, we present a method for selecting control limits for a large family of autocorrelated processes that result in a known average run length in the absence of assignable causes of variation.

A variety of approaches for monitoring autocorrelated processes have been proposed, including the model fitting and EWMA approaches. These approaches have stretched the generally accepted definition of the state of control by accounting for other than chance and assignable causes of variation (i.e. structural cause variation). In Chapters IV and V, we present a new approach for monitoring autocorrelated processes based upon monitoring the capability of the process. Our approach incorporates chance, assignable and structural causes of variation.

2.3 Capability Indices

While maintaining a process in a state of control is of obvious importance, it is also important to determine how well the process is meeting the needs of the customer. Capability indices quantify the performance of the process relative to the needs of the customer. Kane (1986a, 1986b) describes the most widely used capability indices. This section draws heavily upon that work.

2.3.1 Process Potential Index. The process potential index, denoted C_p, measures whether the natural tolerance of the process is within the specification limits for that process. Natural tolerance is arbitrarily defined as six times the process standard deviation, σ. C_p is then defined as the allowable process spread divided by the actual process spread (natural tolerance):

$$C_p = \frac{USL - LSL}{6\sigma}. \tag{45}$$

The general guideline is that a process with a C_p greater than or equal to 1.0 is judged to be capable. Kane recommends a minimum C_p of 1.33 to give "some assurance that a $C_p = 1$ will be possible when additional sources of variance are experienced in production processing." For a normally distributed in-control process, a C_p exactly equal to 1 equates to producing 99.865 percent within the specification limits. Since σ is usually estimated, the process potential can be estimated by $\hat{C}_p$ via

$$\hat{C}_p = \frac{USL - LSL}{6s} \tag{46}$$

where s is the sample standard deviation. Note that the process potential index only relates the process spread to the specification limits and that the location of the process spread relative to the specification limits is not considered. A process with a mean that is much greater than its upper specification limit may produce everything outside of the specification limits while having a *potential* index greater than one.

2.3.2 C_{pk} Index. The second major process index accounts for the location of the process mean relative to the specification limits by combining the upper and lower capability indices. The upper capability index, denoted CPU, is a one-sided version of the process potential index that incorporates the mean and variance of the process along with the upper specification limit. The allowable upper spread is the difference between the upper specification limit and the mean, while the actual upper spread is one-half of the natural tolerance:

$$CPU = \frac{USL - \mu}{3\sigma}. \tag{47}$$

A similar lower capability index is defined as:

$$CPL = \frac{\mu - LSL}{3\sigma}. \tag{48}$$

The C_{pk} index combines the CPL and CPU indices via

$$C_{pk} = \min\{CPL, CPU\} \tag{49}$$

and thus measures the scaled distance between the process mean and the closest specification limit.

2.3.3 The k Index. The final index that Kane describes relates the deviation of the process mean from the midpoint, m, of the specification limits via the equations

$$m = \frac{USL + LSL}{2} \tag{50}$$

and

$$k = \frac{2|m - \mu|}{USL - LSL}. \tag{51}$$

This index isn't very interesting by itself, although it does relate the C_p and C_{pk} indices by the relationship:

$$C_{pk} = C_p(1 - k). \tag{52}$$

When the mean is equidistant from the specification limits, C_{pk} equals C_p. That is, the maximum process potential is realized by centering the process inside of the limits.

2.3.4 The C_{pm} Index. The C_{pm} index is a refinement of the C_{pk}. While the C_{pk} index relates the location of the process mean to the specification limits, the C_{pm} index relates the location of the process mean to the target value. Chan, Cheng and Spiring (1988) propose the index

$$C_{pm} = \frac{USL - LSL}{6\sqrt{\sigma^2 + (\mu - \tau)^2}}. \tag{53}$$

Boyles (Boyles, 1991) discusses the benefits of the C_{pm} index. Specifically, he shows that while the C_{pk} index measures the proportion of items produced inside of the specification limits, it can fail to distinguish between off-target and on-target processes. The C_{pm} index simultaneously monitors process variability and centering on the target.

2.3.5 Estimating Capability Indices. The equations for all three major capability indices (C_p, C_{pk}, C_{pm}) include the process standard deviation, σ. In practice, σ is usually estimated by $\bar{s}$. For the latter two indices, the process mean must also be estimated, generally with $\bar{\bar{x}}$. When a capability index is computed in this way, the result is actually a point estimate of the true capability. Marcucci and Beazley (1988) provide formulas for determining approximate confidence intervals for these estimates when the process is in control.

More recently, Cheng (Cheng, 1994) has proposed a statistical test for capability, as measured by C_p and C_{pm}, assuming the process measurements follow a normal distribution. He states "a process is not considered capable until its capability is proven with an α-risk of making an erroneous decision." He provides the procedure and tables to test C_p and C_{pm} for various sample sizes.

2.3.6 Assumptions Used in Capability Indices. Kane points out that capability indices implicitly assume the process is in control. This point reinforces Wheeler and Chambers conclusion that a process must first be stable before it can be considered to be capable. Kane also points out that, in practice, the capability index relies heavily upon the statistical properties of the estimate of the process standard deviation. We have seen that for an autocorrelated process, this estimate can be grossly biased downward. The impact of this bias is to exaggerate the capability of the process. In addition, the interpretation of the capability indices generally assume the process is normally distributed. When the assumption of normality is violated, Franklin and Wasserman (1992) show that the confidence intervals developed under the assumption of normality for standard capability indices may be overly narrow, exposing the user to the risk of overestimating the true process index.

2.3.7 Tool Wear and Modified Control Charts. It is easy to imagine a process in which the specification limits are quite large relative to the natural tolerance limits. For example, Long and DeCoste (1988) discuss capability studies involving tool wear. They examine a process in which the observations tend to have an upward or downward slope over time due to the effect of tool wear. The mean of the observations is conditioned on time. The process is started with a new tool and has a conditional mean near one specification limit. Over time, the conditional mean shifts towards the other specification limit. A control chart based upon the short-term variation about the conditional mean for this process will have excessively narrow control limits and will therefore generate false alarms as the conditional mean shifts. On the other hand, a control chart based upon the standard deviation of the unconditional distribution as seen over time will be overly wide and may miss significant shifts. Long and DeCoste propose a control chart and capability index that vary with time, similar to using the residuals from a fit model. Quesenberry (1988) also discusses control with regards to a tool wear process.

Montgomery (1991) describes the modified control chart. The modified control chart is intended for monitoring a process in which the natural process tolerance is small compared to the specification limits. It assumes that small shifts in the process mean that do not appreciably affect the fraction of product outside of the specification limits are allowable. In this case, control limits are proposed that are near the specification limits.

2.3.8 Recap. The identification of process capability is recognized as an important tool for quality improvement. Process capability is generally measured by a static capability index, although recent research has recognized that capability can be considered as a dynamic aspect of the process. In Chapter IV, we present the mathematical foundation for dynamically assessing process capability. We use that foundation in Chapter V to develop a method for monitoring a process by monitoring the capability of the process.

2.4 Economic Analysis

All of the control chart techniques described thus far have relied upon 'rule of thumb' design decisions. For example, an $\bar{X}$-chart requires the user to specify the sample size, n, the control limit width factor, k, and how often samples should be taken from the process. A more rational design of a control chart can be accomplished by taking into account the costs associated with the process. For example, a larger sample size may decrease the cost of running a process by detecting any shift in the process sooner, but presumably will increase the cost due to the additional sampling. Economic analysis is concerned with determining the design parameters that minimize the total expected cost of the process.

2.4.1 Shewhart Control Charts. Lorenzen and Vance (1986) provide a general economic design approach for control charts using a renewal reward process. They assume a memoryless process in which the length of time the process stays in control is a negative exponential random variable. They also assume that when the process goes out of control, the mean of the process will shift by a known amount. They cite other research that covers the case for multiple assignable causes. Using the following input variables, they derive the optimal sample size, time between samples and width of the the control limits:

- Time related variables
  - time to sample and chart one item.
  - expected search time for a false alarm.
  - expected time to discover an assignable cause.
  - expected time to repair the process.
- Production decision variables
  - continue production during search?
  - continue production during repair?
- Cost variables
  - quality cost per unit time while in control.
  - quality cost per unit time while not in control.

- cost per false alarm.
- cost to locate and repair the assignable cause.
- fixed cost per sample.
- cost per unit sampled.

2.4.2 Other Control Charts. Goel and Wu (1973) present a procedure for the economic design of a CUSUM chart using the same concepts as Lorenzen and Vance. However, Goel and Wu rely upon a numerical search technique to find the design parameters. Chung (1992) presents a simpler search technique to find the design parameters of a CUSUM control chart. They both assume independent and identically distributed observations and a known mean shift.

2.4.3 The Loss Function. One of the input variables that Lorenzen and Vance use is the quality cost per unit time while either in or out of control. They assume that this value is known. Recently, Taguchi (Taguchi, 1985; Taguchi and Wu, 1980) has put forth the idea that the quality cost is a function of the deviation from the process ideal. By his definition, the ideal target value, τ, for the process is the value which provides the maximum benefit to society as a whole. For a manufacturing process, the ideal target value is generally the same for both the owner of the process and their customers. Any output produced by the process at other than the ideal value will provide less than maximum benefit. It is commonly assumed that the loss of benefits due to the process output being at some point x, where $x \neq \tau$, can be quantified by a nonnegative function, $L(x)$. $L(x)$ is called a loss function or a cost function. Without loss of generality, the ideal loss is assumed to be exactly zero (i.e. $L(\tau) = 0$).

The two most commonly used loss functions are the Kronecker-delta style loss function and the Taguchi loss function. These functions are depicted in Figure 5. The Kronecker loss function penalizes any output outside of the specification limits the same while treating any output inside of the specification limits as having no loss-cost. The Taguchi loss function, on the other hand, assumes that any deviation from the target value will incur a loss.

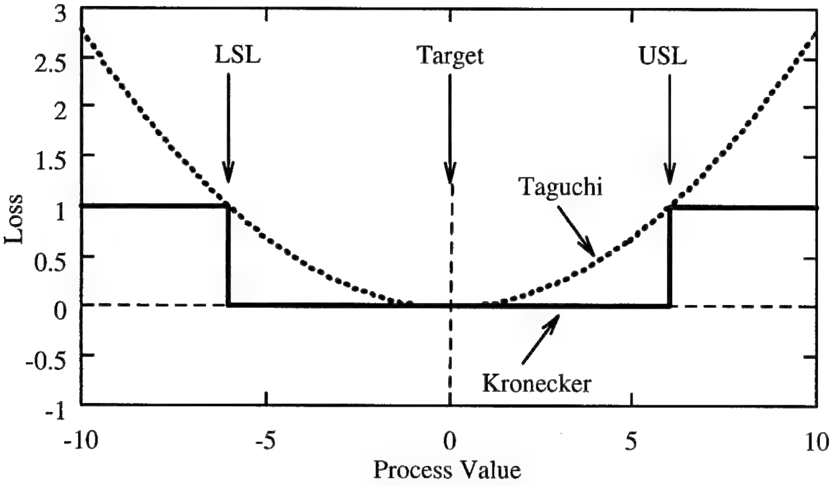


Figure 5. Taguchi and Kronecker delta loss functions.

Taguchi goes further by asserting that any loss function can be approximated by a second order Taylor series expansion, resulting in the equation:

$$L(x) = k(x - \tau)^2 \tag{54}$$

where k is a constant. Neither the Kronecker loss function nor the Taguchi loss function can reasonably be expected to exactly match the true loss function, but both are standard ways of modeling it.

2.5 Chapter Summary.

The primary purpose of this chapter was to present the statistical process control tools in use today and to highlight their limitations when applied to processes that exhibit autocorrelation. All of the standard control charts we discussed (i.e. the $\bar{X}$, X, S, R, CUSUM and EWMA control charts) are specifically designed to test the state of statistical control given independent and identically distributed observations. When applied to observations exhibiting autocorrelation, standard control chart procedures result in unexpectedly reduced average run lengths in the absence of assignable cause variation.

In the next chapter, we present a method for selecting the control limits for an X-chart given an ARMA(1,1) process that provides a specified average run length in the absence of assignable cause variation.

The second purpose of this chapter was to provide background on the use of capability indices. Given the variance of the process, capability indices, such as C_{pk} and C_{pm}, measure its capability. In practice the process variance, and hence, the capability index is estimated for a process that is assumed to be in a state of statistical control. Using statistical techniques, the process capability can be characterized by its mean and a corresponding confidence interval. In Chapter IV we assert that a process can be capable while not necessarily meeting the strict definition of control and that capability can vary over time. For example, both Long and DeCoste's tool wear control chart and the modified control chart recognize that the mean and variance of a process can change over time. In Chapter V, we propose a monitoring system based upon capability.

The final purpose of this chapter was to provide a broader picture of the field of quality improvement so that this research might be kept in perspective. For example, engineering process control approaches take a fundamentally different approach to variation than statistical process control approaches. Both approaches have proven value in different applications. Like the engineering process control and model fitting approaches, the method we propose in Chapter IV for monitoring process capability directly accounts for chance, assignable, and structural causes of variation. We also show in Chapter IV that capability can be considered to measure the economic costs of a process due to deviations from the process target value. This measure is another approximation of the true cost, much like the Kronecker or Taguchi loss functions.

In this and the previous chapter, we explored some of the standard statistical monitoring techniques for quality improvement and identified limitations of those techniques when the process observations exhibit autocorrelation. In the remaining chapters, we develop new methods for monitoring process control and capability in the presence of autocorrelation.

The first new method we develop extends the standard individuals chart to account for autocorrelated observations, specifically for observations arising from ARMA(1,1) processes. In the next chapter, we establish a means for selecting control limits for an ARMA(1,1) process that result in a known in-control average run length.

III. Selecting Control Limits for an ARMA(1,1) Process.

3.1 Introduction

When process measurements are independent and identically distributed, the ability of the standard control charts (i.e. $\bar{X}$, X, R, S, CUSUM and EWMA) to detect shifts in process quality are generally well known. In particular, control limits can be chosen such that a specified in-control average run length is achieved. However, when process measurements exhibit autocorrelation, the standard control charts do not perform as expected. For example, Montgomery and Mastrangelo (1991) report that positive autocorrelation results in an increase in the false alarm rate. This, in turn, implies that the presence of such autocorrelation results in a shorter average run length than would occur for independent and identically distributed process measurements.

The main objective of this chapter is to provide a technique to allow a quality practitioner working with an autocorrelated process to gain the benefits and simplicity of using an X-chart while avoiding the uncertainty caused by the autocorrelation in the process observations. Since the ARMA(1,1) model contains both an autoregressive and moving-average component, it is capable of modelling, at least approximately, the behavior of many real world processes. In this chapter, we develop a method for approximating the average run length for an ARMA(1,1) model with specified upper and lower control limits. We present a table to aid the quality practitioner in choosing appropriate control limits to achieve a desired average run length in the absence of assignable cause variation for an ARMA(1,1) model. Results from this chapter provide the basis for evaluating the capability monitoring approach developed in the following chapters.

Another major objective of this chapter is to better assess the impact of autocorrelation on the average run length. For independent and identically distributed observations, the probability of an observation falling outside of the control limits is a function of the probability density function of the observations, i.e.

$$Pr(x < LCL \text{ or } x > UCL) = 1 - \int_{LCL}^{UCL} f(x)dx \tag{55}$$

where $f(x)$ is the probability density function of the observations. The average run length is a function of the probability of observations falling outside of the control limits, and therefore, depends on this probability density function. However, when the observations exhibit autocorrelation, that probability may change as conditional knowledge about observations is gained over time. In this chapter, we are interested in determining the average run length of an ARMA(1,1) process, which may exhibit autocorrelation. We show that the state of an ARMA(1,1) process can be represented by an ordered pair consisting of the current observation and its underlying error term. The distribution of the state of an ARMA(1,1) process can be described by the joint probability density function of that ordered pair. Then, the probability of an observation from an ARMA(1,1) process falling outside of control limits is a function of that joint probability density function. We further show that the average run length of an ARMA(1,1) process can be expressed as a function of a set of related joint probability density functions. The joint probability density functions are related by each function being successively conditioned on the previous observation falling within the control limits. We show that the average run length can be determined by establishing a relationship between the conditional joint probability function of the state of an ARMA(1,1) process at some time, $t+1$, and the (conditional) joint probability density function of the state of the process at the previous time, t.

This chapter proceeds as follows. In the next section, we provide some background information about the ARMA(1,1) process. After that, we derive the conditional joint probability density function for the state of an ARMA(1,1) process given the joint density function of the previous state. Then, we develop a recursive relationship which identifies the changes in the probability density function over time by incorporating conditional information gained about the process through time. We use that recursive relationship is used to derive the average run length for an ARMA(1,1) process with specified control limits. We also present some illustrative examples, followed by tables which summarize the key results. For easy reference, the notation used in this chapter is summarized in Appendix A.

3.2 Background on the ARMA(1,1) Model.

We present some background, notation and important features of the ARMA(1,1) model in this section. The ARMA(1,1) model is important because many real world processes can, at least approximately, be modelled with the ARMA(1,1) model. An ARMA(1,1) process is characterized by the relationship

$$x_{t+1} = \xi + \phi x_t - \theta \epsilon_t + \epsilon_{t+1} \tag{56}$$

where x_t is the process observation at time t, ϵ_t is the zero mean random error at time t, ξ is a constant that adjusts for the mean of the process, ϕ is the autoregressive parameter and θ is the moving average parameter. For simplicity and without loss of generality, we further assume that the process is centered at zero and, thus, is characterized by

$$x_{t+1} = \phi x_t - \theta \epsilon_t + \epsilon_{t+1}. \tag{57}$$

From this equation it is clear that the state at some time, t, of a zero mean ARMA(1,1) process can be specified by the ordered pair (x_t, ϵ_t): the observation and underlying error at time t. All future states can be expressed as a function of the current state and future unknown errors by recursively applying equation 57. For instance,

$$\begin{aligned} x_{t+2} &= \phi^2 x_t - \phi\theta\epsilon_t + \phi\epsilon_{t+1} - \theta\epsilon_{t+1} + \epsilon_{t+2}, \\ x_{t+3} &= \phi^3 x_t - \phi^2\theta\epsilon_t + \phi^2\epsilon_{t+1} - \phi\theta\epsilon_{t+1} + \phi\epsilon_{t+2} - \theta\epsilon_{t+2} + \epsilon_{t+3}, \end{aligned} \tag{58}$$

and, in general, for $k \geq 1$,

$$x_{t+k} = \phi^k x_t - \phi^{k-1}\theta\epsilon_t + \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)\epsilon_{t+i} + \epsilon_{t+k}. \tag{59}$$

In order to ensure that the process has a stable mean, we will restrict ourselves to stationary ARMA(1,1) processes. Stationarity implies that $-1 < \phi < 1$ (Box and Jenkins, 1976). From equation 59, it should be clear that the influence of the current state on a future state decreases as the time until the future state, k, becomes large. In the limiting case, we have

$$\lim_{k\to\infty} \phi^k = 0 \tag{60}$$

so, when x_t and ϵ_t are finite,

$$\begin{aligned} \lim_{k\to\infty} x_{t+k} &= \lim_{k\to\infty} \phi^k x_t - \lim_{k\to\infty} \phi^{k-1}\epsilon_t + \lim_{k\to\infty} \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)\epsilon_{t+i} + \lim_{k\to\infty} \epsilon_{t+k} \\ &= \lim_{k\to\infty} \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)\epsilon_{t+i} + \lim_{k\to\infty} \epsilon_{t+k}. \end{aligned} \tag{61}$$

The reduced contribution of the current state can also be seen by considering the conditional mean and variance of the observation to be made k time steps into the future:

$$\begin{aligned} E[X_{t+k}|x_t, \epsilon_t] &= E[\phi^k X_t - \phi^{k-1}\theta\mathcal{E}_t + \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)\mathcal{E}_{t+i} + \mathcal{E}_{t+k}|x_t, \epsilon_t] \\ &= \phi^k E[X_t|x_t, \epsilon_t] - \phi^{k-1}\theta E[\mathcal{E}_t|x_t, \epsilon_t] \\ &= \phi^k\ x_t - \phi^{k-1}\theta\epsilon_t \end{aligned} \tag{62}$$

and

$$\begin{aligned}
Var[X_{t+k}|x_t, \epsilon_t] &= Var[\phi^k X_t - \phi^{k-i}\theta\mathcal{E}_t + \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)\mathcal{E}_{t+i} + \mathcal{E}_{t+k}|x_t, \epsilon_t] \\
&= \phi^{2(k)}Var[X_t|x_t, \epsilon_t] + \phi^{2(k-i)}\theta^2 Var[\mathcal{E}_t|x_t, \epsilon_t] + \\
&\qquad \sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)^2 Var[\mathcal{E}_{t+i}|x_t, \epsilon_t] + Var[\mathcal{E}_{t+k}|x_t, \epsilon_t] \\
&= [\sum_{i=1}^{k-1}(\phi^{k-i} - \phi^{k-i-1}\theta)^2 + 1]\ \sigma_\epsilon^2. \qquad (63)
\end{aligned}$$

Both the expected value and the variance of future observations are clearly influenced by the current state. However, the influence of the current state drops rapidly even when $|\phi|$ is close to one. For example, suppose we have an AR(1) process with $\phi = .95$ and let x_t and ϵ_t be samples from the random variables X_t and $\mathcal{E}_t$, respectively. From Montgomery (1990) we know that the unconditional variance of X_t, denoted σ_x^2, is related to the variance of the errors, denoted σ_ϵ^2, via

$$\sigma_x^2 = \sigma_\epsilon^2 \frac{1}{1-\phi^2} = 10.26\ \sigma_\epsilon^2. \qquad (64)$$

The ratio of the variance of the k-step ahead observation to the unconditional process observation, $Var(X_{t+k}|x_t, \epsilon_t)/Var(X_t)$, for this AR(1) process is depicted in the top half of Figure 6. The ratio rapidly approaches its limit of 1. In addition, the ratio of the conditional expected value of the future observation to the current observation is depicted in the bottom half of the figure. That ratio approaches zero as the number of time steps into the future increases. In other words, the conditional expected value of the process observations approaches the unconditional mean of the process, and, the conditional variance approaches the unconditional variance.

The decreasing influence of conditional information extends to the full ARMA family of models. In general, any ARMA process can be represented by the general linear filter

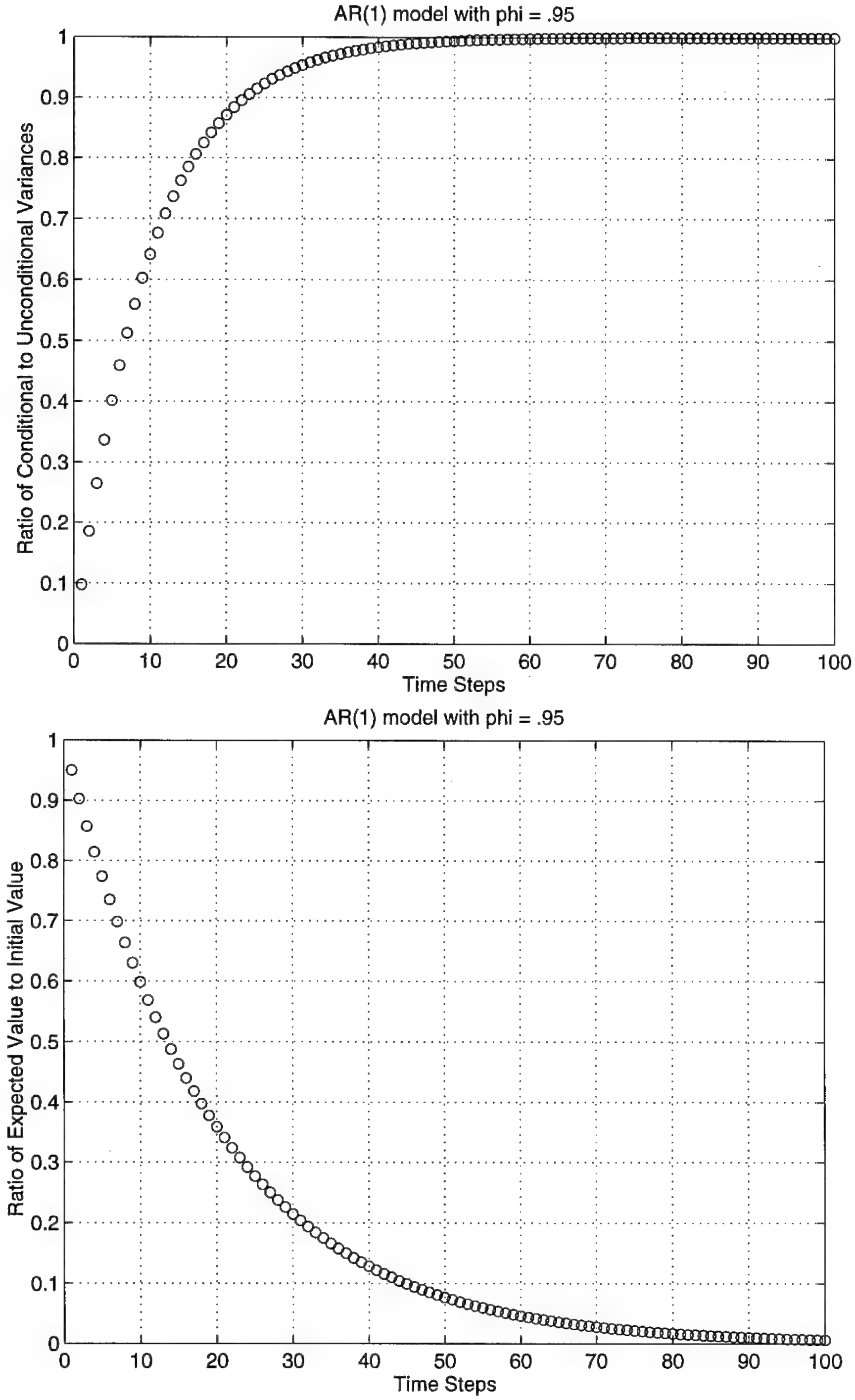


Figure 6. Convergence of the conditional variance to the unconditional variance and the conditional mean to zero for an AR(1) process with $\phi = 0.95$.

(Montgomery, 1991)

$$x_t = \mu + \sum_{i=0}^{\infty} \psi_i \epsilon_{t-i} \tag{65}$$

where μ is the mean of the process, the sequence of weights, $\{\psi_i\}$, are constant, and the random shocks, $\{\epsilon_i\}$, which drive the system are independent and identically distributed random variables with mean 0 and variance σ_ϵ^2. When the process is stationary, the sequence $\{\psi_i\}$ is either finite or infinite and convergent. A finite sequence implies that the influence of the current state on future states disappears after a specified number of time steps, while a convergent sequence implies that the influence of the current state decreases over time. Thus, for any stationary ARMA process, the impact of previous process states on current and future states diminishes over time. Later on, we will use this property to approximate the average run length of an ARMA(1,1) process with fixed control limits.

3.3 Defining the State of an ARMA(1,1) Process and Deriving its Probability Density Function.

In this section, we present the mathematical foundation for approximating the average run length for an ARMA(1,1) process. First, we define the state of an ARMA(1,1) process in terms of the process observation and underlying error. Then, we derive the joint probability density function of the next state of the process given the joint probability density function of the current state.

Let the process observation and underlying error at time t, denoted x_t and ϵ_t, be samples from the random variables X_t and $\mathcal{E}_t$, respectively. The process state at time $t+1$ can be described by these two random variables where

$$X_{t+1} = \phi X_t - \theta \mathcal{E}_t + \mathcal{E}_{t+1} \tag{66}$$

and $\mathcal{E}_{t+1}$ represents an error distribution with zero mean and finite variance. Define $f(x_t, \epsilon_t)$ as the joint probability density function of the state of the process at time t. Further denote the marginal probability density function of the process observations at time t as $f^*(x_t)$. It is clear that X_{t+1} and $\mathcal{E}_{t+1}$ depend only on $(X_t, \mathcal{E}_t,\ \mathcal{E}_{t+1})$ where, by definition, $\mathcal{E}_{t+1}$ is independent of X_t and $\mathcal{E}_t$. Thus, the joint probability density function of $(X_t, \mathcal{E}_t,\ \mathcal{E}_{t+1})$ is given by $f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1})$ where Φ' is the probability density function of the error distribution. The joint probability density function of $(X_{t+1},\ \mathcal{E}_{t+1})$ will be developed from the joint probability density function of $(X_t, \mathcal{E}_t,\ \mathcal{E}_{t+1})$. Let $G(y_1, y_2)$ denote the cumulative density function of $(X_{t+1},\ \mathcal{E}_{t+1})$ evaluated at the point $(y_1,\ y_2)$. We will use y_1 and y_2 instead of x_{t+1} and ϵ_{t+1} for now to avoid confusion with the limits of integration. Then, we can write (DeGroot, 1989)

$$G(y_1, y_2) = \int\int\int_{A_y} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) dx_t d\epsilon_t d\epsilon_{t+1} \tag{67}$$

where A_y is defined as the subset of R^3 containing all $(x_t, \epsilon_t, \epsilon_{t+1})$ such that

$$\begin{aligned} \phi x_t - \theta\epsilon_t + \epsilon_{t+1} &\leq y_1 \\ \epsilon_{t+1} &\leq y_2. \end{aligned} \tag{68}$$

Note that the order of integration will be determined by the definition of A_y. Three cases must now be considered which depend on the autoregressive and moving average parameters, ϕ and θ respectively. The first case, with $\phi = 0$ and $\theta = 0$, considers a simple processes in which the observations are independent and identically distributed. The second case, with $\theta \neq 0$, considers both mixed ARMA(1,1) processes and pure MA(1) processes. The third case, with $\phi \neq 0$ and $\theta = 0$, considers pure AR(1) processes.

3.3.1 Case 1: $\phi = 0$ and $\theta = 0$. When both ϕ and θ are equal to zero, the observations are simply the independent and identically distributed errors. Recall from Chapters I and II that this is the base model that much of the SPC techniques were developed for. Since $X_{t+1} = \mathcal{E}_{t+1}$ for this case (see equation 66), the joint probability density function is a degenerate case. The marginal probability density functions of X_{t+1} and $\mathcal{E}_{t+1}$ are the same and will be denoted as $g^*(y)$. We will immediately proceed to derive

$g^*(y)$ by first deriving the marginal cumulative density function, $G^*(y)$. The region A_y is constrained by $\epsilon_{t+1} \leq y_1$ from equation 68. Therefore,

$$\begin{aligned} G^*(y) &= \int_{-\infty}^{y} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_t, \epsilon_t) \Phi'(\epsilon_{t+1}) dx_t d\epsilon_t d\epsilon_{t+1} \\ &= \int_{-\infty}^{y} \Phi'(\epsilon_{t+1}) d\epsilon_{t+1} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x_t, \epsilon_t) dx_t d\epsilon_t \\ &= \Phi(y), \end{aligned} \tag{69}$$

and, as expected, the probability density function is

$$g^*(y) = \Phi'(y). \tag{70}$$

3.3.2 Case 2: $\theta \neq 0$. This case includes both the mixed ARMA(1,1) model and the pure MA(1) model. First, consider the sub-case in which $\theta < 0$. To identify the region A_y, allow x_t to take on any value and then apply the second constraint in equation 68 to limit ϵ_{t+1} to be less than y_2. Then from the first constraint, $\epsilon_t \leq (\phi x_t + \epsilon_{t+1} - y_1)/\theta$. Therefore, from equation 67

$$G(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \int_{-\infty}^{(\phi x_t + \epsilon_{t+1} - y_1)/\theta} f(x_t, \epsilon_t) \Phi'(\epsilon_{t+1}) d\epsilon_t d\epsilon_{t+1} dx_t. \tag{71}$$

The joint probability density function of $(X_{t+1}, \mathcal{E}_{t+1})$ can then be computed by taking the derivative of the cumulative distribution function with respect to y_1 and y_2. A sufficient condition for differentiating under the first two integral signs is for the joint density function f to be continuous almost everywhere and to decay to zero in the limit of x. That is,

$$\begin{aligned} \lim_{x_t \to \infty} f(x_t, \epsilon_t) &= 0 \quad \forall\, \epsilon_t \\ \lim_{x_t \to -\infty} f(x_t, \epsilon_t) &= 0 \quad \forall\, \epsilon_t. \end{aligned} \tag{72}$$

Given that condition,

$$
\begin{aligned}
g(y_1, y_2) &= \frac{\partial}{\partial y_2}\frac{\partial}{\partial y_1} G(y_1, y_2) \\
&= \frac{\partial}{\partial y_2}\frac{\partial}{\partial y_1} \int_{-\infty}^{\infty}\int_{-\infty}^{y_2}\int_{-\infty}^{(\phi x_t + \epsilon_{t+1} - y_1)/\theta} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) d\epsilon_t d\epsilon_{t+1} dx_t \\
&= \frac{\partial}{\partial y_2} \int_{-\infty}^{\infty}\int_{-\infty}^{y_2} \frac{\partial}{\partial y_1}\int_{-\infty}^{(\phi x_t + \epsilon_{t+1} - y_1)/\theta} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) d\epsilon_t d\epsilon_{t+1} dx_t \qquad (73)
\end{aligned}
$$

And, by applying Liebnitz's Rule for differentiating under the integral,

$$
\begin{aligned}
g(y_1, y_2) &= \frac{\partial}{\partial y_2} \int_{-\infty}^{\infty}\int_{-\infty}^{y_2} \frac{\partial(\phi x_t + \epsilon_{t+1} - y_1)/\theta)}{\partial y_1} f(x_t, (\phi x_t + \epsilon_{t+1} - y_1)/\theta)\Phi'(\epsilon_{t+1}) d\epsilon_{t+1} dx_t \\
&= \frac{\partial}{\partial y_2} \int_{-\infty}^{\infty}\int_{-\infty}^{y_2} \frac{-1}{\theta} f(x_t, (\phi x_t + \epsilon_{t+1} - y_1)/\theta)\Phi'(\epsilon_{t+1}) d\epsilon_{t+1} dx_t \\
&= \int_{-\infty}^{\infty} \frac{\partial}{\partial y_2} \int_{-\infty}^{y_2} \frac{-1}{\theta} f(x_t, (\phi x_t + \epsilon_{t+1} - y_1)/\theta)\Phi'(\epsilon_{t+1}) d\epsilon_{t+1} dx_t \\
&= \int_{-\infty}^{\infty} \frac{-1}{\theta} f(x_t, (\phi x_t + y_2 - y_1)/\theta)\Phi'(y_2) dx_t \\
&= \frac{-1}{\theta}\Phi'(y_2) \int_{-\infty}^{\infty} f(x_t, (\phi x_t + y_2 - y_1)/\theta) dx_t. \qquad (74)
\end{aligned}
$$

The second sub-case, for $\theta > 0$, follows the same line of development but includes a sign change to yield

$$
g(y_1, y_2) = \frac{1}{\theta}\Phi'(y_2) \int_{-\infty}^{\infty} f(x_t, (\phi x_t + y_2 - y_1)/\theta) dx_t. \qquad (75)
$$

The two sub-cases can therefore be combined into the single case when $\theta \neq 0$ via

$$
g(y_1, y_2) = |\frac{1}{\theta}|\Phi'(y_2) \int_{-\infty}^{\infty} f(x_t, (\phi x_t + y_2 - y_1)/\theta) dx_t. \qquad (76)
$$

y_1 and y_2 can now be replaced by x_{t+1} and ϵ_{t+1} to yield

$$g(x_{t+1}, \epsilon_{t+1}) = |\frac{1}{\theta}|\Phi'(\epsilon_{t+1}) \int_{-\infty}^{\infty} f(x_t, (\phi x_t + \epsilon_{t+1} - x_{t+1})/\theta) dx_t, \tag{77}$$

or,

$$g(x_t, \epsilon_t) = |\frac{1}{\theta}|\Phi'(\epsilon_t) \int_{-\infty}^{\infty} f(x_{t-1}, (\phi x_{t-1} + \epsilon_t - x_t)/\theta) dx_{t-1}, \tag{78}$$

or,

$$g(x_t, \epsilon_t) = |\frac{1}{\theta}|\Phi'(\epsilon_t) \int_{-\infty}^{\infty} f(x_{t-1}, \epsilon_{t-1}) dx_{t-1}. \tag{79}$$

3.3.3 Case 3: $\phi \neq 0$ and $\theta = 0$. The final case to be considered is when $\theta = 0$ and $\phi \neq 0$. This is the AR(1) case. First, consider the sub-case in which $\phi > 0$. The region A_y can be identified by allowing ϵ_t to take on any value and using the second constraint in equation 68 to limit ϵ_{t+1} to be less than y_2. Also, from the first constraint, $x_t \leq (y_1 - \epsilon_{t+1})/\phi$. Then from equation 67

$$G(y_1, y_2) = \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \int_{-\infty}^{(y_1-\epsilon_{t+1})/\phi} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) dx_t d\epsilon_{t+1} d\epsilon_t. \tag{80}$$

The joint probability density function of (y_1, y_2) can then be computed by taking the derivative of the cumulative density function with respect to y_1 and y_2:

$$\begin{aligned} g(y_1, y_2) &= \frac{\partial}{\partial y_2}\frac{\partial}{\partial y_1} G(y_1, y_2) \\ &= \frac{\partial}{\partial y_2}\frac{\partial}{\partial y_1} \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \int_{-\infty}^{(y_1-\epsilon_{t+1})/\phi} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) dx_t d\epsilon_{t+1} d\epsilon_t \\ &= \frac{\partial}{\partial y_2} \int_{-\infty}^{\infty} \int_{-\infty}^{y_2} \frac{\partial}{\partial y_1} \int_{-\infty}^{(y_1-\epsilon_{t+1})/\phi} f(x_t, \epsilon_t)\Phi'(\epsilon_{t+1}) dx_t d\epsilon_{t+1} d\epsilon_t. \end{aligned} \tag{81}$$

And, by applying Liebnitz's Rule for differentiating under the integral,

$$\begin{aligned}
g(y_1, y_2) &= \frac{\partial}{\partial y_2}\int_{-\infty}^{\infty}\int_{-\infty}^{y_2}\frac{\partial (y_1-\epsilon_{t+1})/\phi}{\partial y_1} f((y_1-\epsilon_{t+1})/\phi, \epsilon_t)\Phi'(\epsilon_{t+1})d\epsilon_{t+1}d\epsilon_t \\
&= \frac{\partial}{\partial y_2}\int_{-\infty}^{\infty}\int_{-\infty}^{y_2}\frac{1}{\phi} f((y_1-\epsilon_{t+1})/\phi, \epsilon_t)\Phi'(\epsilon_{t+1})d\epsilon_{t+1}d\epsilon_t \\
&= \int_{-\infty}^{\infty}\frac{\partial}{\partial y_2}\int_{-\infty}^{y_2}\frac{1}{\phi} f((y_1-\epsilon_{t+1})/\phi, \epsilon_t)\Phi'(\epsilon_{t+1})d\epsilon_{t+1}d\epsilon_t \\
&= \int_{-\infty}^{\infty}\frac{1}{\phi} f((y_1-y_2)/\phi, \epsilon_t)\Phi'(y_2)d\epsilon_t \\
&= \frac{1}{\phi}\Phi'(y_2)\int_{-\infty}^{\infty} f((y_1-y_2)/\phi, \epsilon_t)d\epsilon_t \\
&= \frac{1}{\phi}\Phi'(y_2)f^*((y_1-y_2)/\phi). \qquad (82)
\end{aligned}$$

Similar to the previous case, a sign change occurs when $\phi < 0$ so that the general case becomes

$$g(y_1, y_2) = |\frac{1}{\phi}|\Phi'(y_2)f^*((y_1-y_2)/\phi). \qquad (83)$$

Since future observations from an AR(1) process depend only on the current observation and not the current error, it is useful to consider the marginal distribution of the observations. Using the definition of marginal distribution,

$$\begin{aligned}
g^*(y_1) &= \int_{-\infty}^{\infty} g(y_1, y_2)dy_2 \\
&= \int_{-\infty}^{\infty} |\frac{1}{\phi}|\Phi'(y_2)f^*((y_1-y_2)/\phi)dy_2. \qquad (84)
\end{aligned}$$

Later, it will be convenient for numerical evaluation of this expression to transform the variables such that the function f is directly evaluated over the range of the integral rather than being evaluated as a function of the range. To that end, let $Q = (y_1 - y_2)/|\phi|$. Then

$y_2 = y_1 - |\phi|Q$, $dy_2 = -|\phi|dQ$, and

$$
\begin{aligned}
g^*(y_1) &= \int_{Q=\infty}^{Q=-\infty} -\frac{1}{|\phi|}\Phi'(y_2)f^*((y_1-y_2)/\phi)|\phi|dQ \\
&= \int_{-\infty}^{\infty} \Phi'(y_1 - \phi Q)f^*(Q)dQ. \qquad (85)
\end{aligned}
$$

A better understanding may be gained by now substituting x_{t+1} for y_1 and replacing Q with x_t, so that the previous equation becomes

$$g^*(x_{t+1}) = \int_{-\infty}^{\infty} \Phi'(x_{t+1} - \phi x_t)f^*(x_t)dx_t, \qquad (86)$$

or,

$$g^*(x_t) = \int_{-\infty}^{\infty} \Phi'(\epsilon_t)f^*(x_{t-1})dx_{t-1}. \qquad (87)$$

3.4 Recursive Representation of Density Function for the State of an ARMA(1,1) Model.

In the previous section, we showed how the unconditional probability density function for the next state of an ARMA(1,1) process can be expressed as a function of the probability density function of the current state. The next step in determining the average run length for an ARMA(1,1) process with fixed control limits is to develop a recursive representation for the state of the process which also incorporates information gained about process observations over time. In the previous section, we developed formulas to express the density function of the next ARMA(1,1) state in terms of the current density function. In this section, we use those results to develop the conditional density function of the next ARMA(1,1) state given that the next observation is between predetermined lower and upper control limits.

Define $f_n(x_t, \epsilon_t)$ to be the conditional joint probability density function of the state of an ARMA(1,1) process at time t given that the n most recent observations (i.e., those observed between time $t-n+1$ and t) have been within the specified control limits. That is, given upper and lower control limits, denoted LCL and UCL respectively,

$$f_n(x_t, \epsilon_t) = f(x_t, \epsilon_t | LCL \leq x_t \leq UCL, \ldots, LCL \leq x_{t-n+1} \leq UCL), \qquad (88)$$

and

$$f_{n+1}(x_t, \epsilon_t) = f(x_t, \epsilon_t | LCL \leq x_t \leq UCL, \ldots, LCL \leq x_{t-n} \leq UCL). \qquad (89)$$

$f_0(x_t, \epsilon_t)$ is interpreted as the unconditional joint probability density function of the state of an ARMA(1,1) process at time t. Further, define $g_{n+1}(x_t, \epsilon_t)$ as the conditional joint density function of the state of an ARMA(1,1) process at time t given that the *previous* n observations (i.e., those observed between times $t - n$ and $t - 1$) were between the control limits, via

$$g_{n+1}(x_t, \epsilon_t) = f(x_t, \epsilon_t | LCL \leq x_{t-1} \leq UCL, \ldots, LCL \leq x_{t-n} \leq UCL). \qquad (90)$$

Thus, $f_{n+1}(x_t, \epsilon_t)$ is related to $g_{n+1}(x_t, \epsilon_t)$ by incorporating the conditional information that the observation at time t is between LCL and UCL. This relationship can be written as

$$\begin{aligned} g_{n+1}(x_t, \epsilon_t | LCL \leq x_t \leq UCL) &= f(x_t, \epsilon_t | LCL \leq x_t \leq UCL \ \text{ AND} \\ &\qquad LCL \leq x_{t-1} \leq UCL, \ldots, LCL \leq x_{t-n} \leq UCL) \\ &= f(x_t, \epsilon_t | LCL \leq x_t \leq UCL, \ldots, LCL \leq x_{t-n} \leq UCL) \\ &= f_{n+1}(x_t, \epsilon_t) \qquad (91) \end{aligned}$$

which can be rewritten as

$$f_{n+1}(x_t, \epsilon_t) = \begin{cases} \dfrac{g_{n+1}(x_t, \epsilon_t)}{Pr(LCL \leq X_t \leq UCL)} & \text{if } \ LCL \leq x_t \leq UCL \\ 0 & \text{otherwise} \end{cases} \qquad (92)$$

or

$$f_{n+1}(x_t,\epsilon_t) = \begin{cases} \dfrac{g_{n+1}(x_t,\epsilon_t)}{\int_{LCL}^{UCL}\int_{-\infty}^{\infty} g_{n+1}(x_t,\epsilon_t)d\epsilon_t dx_t} & \text{if } LCL \le x_t \le UCL \\ 0 & \text{otherwise.} \end{cases} \tag{93}$$

The relationship for the marginal distributions can be similarly written as

$$f^*_{n+1}(x_t) = \begin{cases} \dfrac{g^*_{n+1}(x_t)}{\int_{LCL}^{UCL} g^*_{n+1}(x_t)dx_t} & \text{if } LCL \le x_t \le UCL \\ 0 & \text{otherwise.} \end{cases} \tag{94}$$

A recursive relationship for each of the three cases from the previous section can now be identified.

3.4.1 Case 1: $\phi = 0$ and $\theta = 0$.

$$\begin{aligned} g^*_{n+1}(x_t) &= \Phi'(x_t) \\ f^*_{n+1}(x_t) &= \begin{cases} \dfrac{\Phi'(x_t)}{\Phi(UCL) - \Phi(LCL)} & \text{if } LCL \le x_t \le UCL \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{95}$$

3.4.2 Case 2: $\theta \neq 0$.

$$\begin{aligned} g_{n+1}(x_t,\epsilon_t) &= |\frac{1}{\theta}|\Phi'(\epsilon_t)\int_{-\infty}^{\infty} f_n(x_{t-1},(x_t - \phi x_{t-1} - \epsilon_t)/\theta)dx_{t-1} \\ f_{n+1}(x_t,\epsilon_t) &= \begin{cases} \dfrac{g_{n+1}(x_t,\epsilon_t)}{\int_{LCL}^{UCL}\int_{-\infty}^{\infty} g_{n+1}(x_t,\epsilon_t)d\epsilon_t dx_t} & \text{if } LCL \le x_t \le UCL \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{96}$$

3.4.3 Case 3: $\phi \neq 0$ and $\theta = 0$.

$$\begin{aligned} g^*_{n+1}(x_t) &= \int_{-\infty}^{\infty} \Phi'(x_t - \phi Q) f^*_n(Q) dQ \\ f^*_{n+1}(x_t) &= \begin{cases} \dfrac{g^*_{n+1}(x_t)}{\int_{LCL}^{UCL} g^*_{n+1}(x_t) dx_t} & \text{if } LCL \leq x_t \leq UCL \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{97}$$

It is not difficult to see in equation 96 that $f_{n+1}(x_t, \epsilon_t)$ can be expressed in terms of $f_n(x_{t-1}, \epsilon_{t-1})$ rather than $g_{n+1}(x_t, \epsilon_t)$ by combining equations 96 and 79. Similarly, $f^*_{n+1}(x_t)$ can be expressed in terms of $f^*_n(x_{t-1})$ in equation 97. However, the functions $g_{n+1}(x_t, \epsilon_t)$ and $g^*_{n+1}(x_t)$ provide both an intuitive intermediate step and a convenient computational breakpoint for numerically evaluating the equations. As an example, Figure 7 on page 61 depicts a series of marginal probability density functions for an AR(1) process with high autocorrelation ($\phi = 0.9$) and narrow control limits ($\pm 2\sigma_x$). The sequence begins in the upper right with the unconditional probability density function of X_t, $f^*_0(x_t)$. The standard deviation of the measurements from this process is 2.29 and its natural tolerance limit, defined as $\pm 3\sigma_x$, is therefore -6.87 to 6.87. Note that since $g^*_n(x_t)$ is defined as the conditional distribution of X_t given that the $n-1$ previous observations were within the control limits, $g^*_1(x_t)$ actually does not incorporate any conditional information and so $g^*_1(x_t)$ equals $f^*_0(x_t)$.

Figure 7 also provides visual evidence of the convergence of both f_n and g_n as n gets large. That is, as n becomes large, the probability distribution function f_n approaches some limiting function. We will use this property later on to approximate an average run length. Looking at the sequence $\{f_0, f_1, f_2, f_3\}$, we can see that the incremental changes decrease significantly in only a few iterations. While this example is purposely exaggerated for visual appeal by the choice of very narrow control limits, it is our experience for that a similar pattern of convergence can be expected for wider control limits and for a variety of ARMA(1,1) processes. The incremental changes are most significant near the control

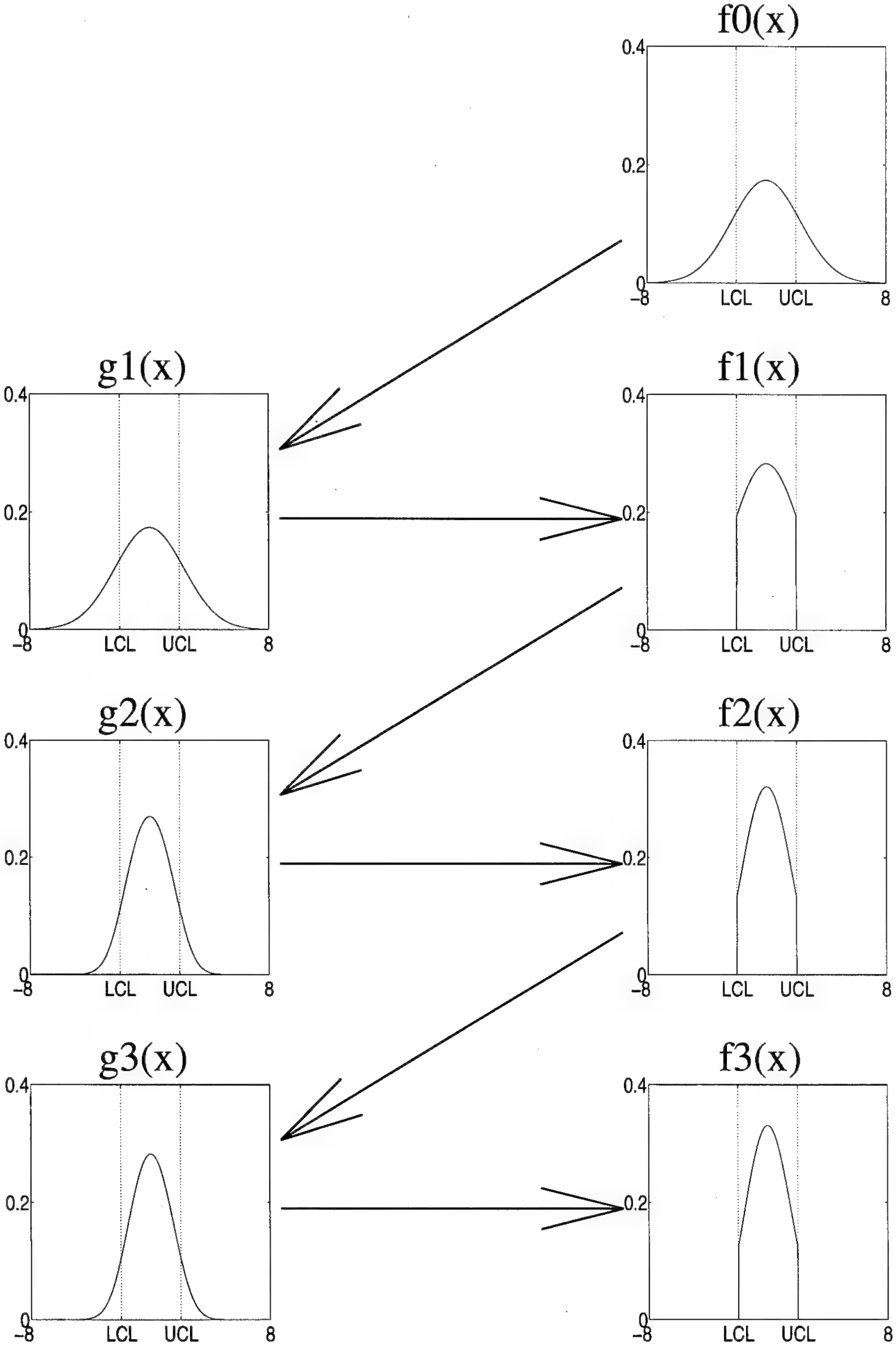


Figure 7. Probability Density Functions for the state of an AR(1) process with standard normal errors, $\phi = .9$, LCL $= -2\sigma_x$, and UCL $= 2\sigma_x$.

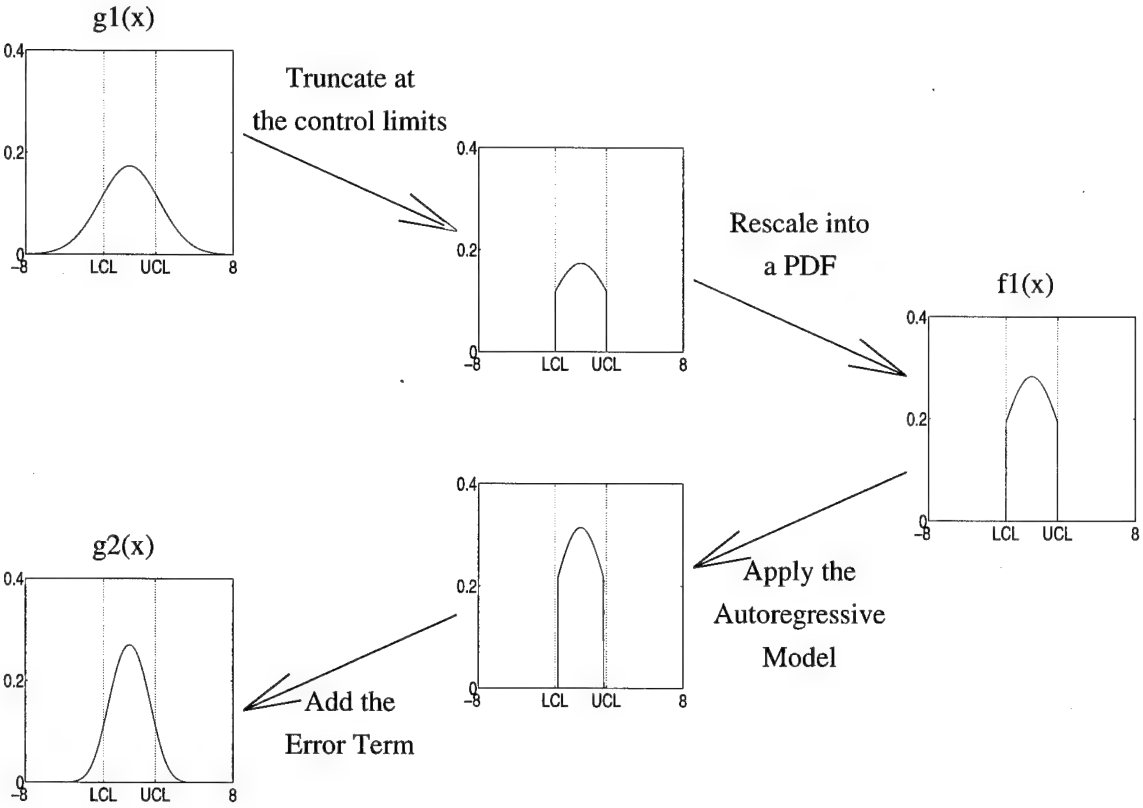


Figure 8. Graphical illustration of the recursive relationship between f and g.

limits, thus, when control limits are chosen in the tails of the distributions, the changes are simply difficult to show graphically.

Figure 8 further breaks down the mechanics underlying one iteration. The transition from $g_1^*(x_t)$ to $f_1^*(x_t)$ can be graphically recreated in two steps. First, the tails of $g_1^*(x_t)$ are truncated at the control limits to incorporate the conditional information that $LCL < x_1 < UCL$. Second, the truncated curve is rescaled such that the area under it equals one, turning it into the probability density function $f_1^*(x_t)$. The transition from $f_1^*(x_t)$ to $g_2^*(x_t)$ can also be envisioned as first applying the autoregressive equation and then adding in the normal error. Similar physical constructs can be envisioned for a full ARMA(1,1) model, albeit in three dimensions rather than just two.

3.5 Determining the Average Run Length

We will proceed to show that the recursive relationships developed in the previous section can be used to compute an average run length for an ARMA(1,1) process with given control limits. However, Three major problems arise in trying to do so. First, the average run length until an out of control signal for an ARMA(1,1) process depends upon the conditional false alarm rate, which changes as the number of in-control observations increases. Second, the initial density functions, $f_0(x_t, \epsilon_t)$ and $f_0^*(x_t)$, are not necessarily known. Third, although the relationships are concisely stated in equations 95, 96 and 97, they are not directly solvable. Some numerical method must be used to approximate the exact solution. Each of these points will be discussed in turn.

3.5.1 Average Run Length for ARMA(1,1) Models. The first problem is how to use the conditional joint probability density functions to compute an average run length. For the independent and identically distributed model, the average run length is quite simple to calculate. In the absence of any assignable cause variation, the false alarm rate at every time is equal to the probability of a type I error, denoted α, or, the probability of an observation being outside of the control limits. For the independent and identically distributed case

$$\alpha = 1 - \Phi(UCL) + \Phi(LCL). \tag{98}$$

For this case, the average run length, starting at any time, can be computed via the infinite sum

$$\begin{aligned} ARL &= \sum_{i=1}^{\infty} i \; Pr(\text{Run length} = i) \\ &= \sum_{i=1}^{\infty} i \; \alpha \; (1-\alpha)^{i-1} \end{aligned} \tag{99}$$

which simplifies to $ARL = 1/\alpha$.

For more complex ARMA(1,1) models, in which either $\phi \neq 0$ or $\theta \neq 0$, structural cause variation will be present. This variation influences the joint probability density functions, g_n and f_n. In the absence of assignable cause variation, the probability of the observation at time $t+n$ falling outside of the control limits given that the observations between time t and $t+n-1$ were within the control limits, denoted α_{t+n}, is a function of the joint probability density function, $g_n(x_{t+n}, \epsilon_{t+n})$, via the equation

$$\alpha_{t+n} = 1 - \int_{LCL}^{UCL} \int_{-\infty}^{\infty} g_n(x_{t+n}, \epsilon_{t+n}) d\epsilon_{t+n} dx_{t+n}. \tag{100}$$

Clearly, α_{t+n} is conditioned upon the observations between time t and $t+n-1$ falling within the control limits. The run length is defined as the number of observations until the first observation which falls outside of the control limits. The probability that the first $i-1$ observations after t are all within the control limits and the ith observation is outside of the control limits can be expressed via

$$Pr(\text{Run length} = i) = \alpha_{t+i} \prod_{j=1}^{i-1} (1 - \alpha_{t+j}). \tag{101}$$

Then, the average run length starting at time t, denoted ARL_t, can be computed via

$$ARL_t = \sum_{i=1}^{\infty} i \; Pr(\text{Run length} = i) \tag{102}$$

or

$$ARL_t = \sum_{i=1}^{\infty} i \; \alpha_{t+i} \; \prod_{j=1}^{i-1} (1 - \alpha_{t+j}). \tag{103}$$

The preceding equation is similar to the independent case (equation 99). However, since the α_{t+i}'s are not generally constant for the dependent cases, ARL_t is not, in general, equal to $1/\alpha_t$.

A key observation will greatly clarify this situation. For all stationary ARMA(1,1) models, past observations have increasingly smaller impacts over time on the present observations. That is, the change in the marginal distribution of X due to conditional knowledge about a particular past observation diminishes as the time since that observation grows, or,

$$\lim_{n\to\infty} g^*(x_{t+n}|x_t) = g^*(x_{t+n}). \tag{104}$$

Similarly,

$$\lim_{n\to\infty} g_n^*(x_{t+n}) = \lim_{n\to\infty} g_{n+m}^*(x_{t+n+m}) \qquad \forall\, m > 0. \tag{105}$$

which, in turn, implies that

$$\lim_{n\to\infty} \alpha_n = \lim_{n\to\infty} \alpha_{n+m} \qquad \forall\, m > 0. \tag{106}$$

Based on computational experience (to be discussed later in this chapter), we have found that the distribution of the observations can be reasonably approximated by a limiting distribution for n as small as 30 for stationary ARMA(1,1) models. Thus, the false alarm rate any time after 30 in-control observations can be reasonably approximated by the false alarm rate after exactly 30 in-control observations. Therefore, the conditional average run length starting at time t given that 30 or more observations immediately prior to time t were within the control limits, denoted as $ARL_{t|30}$, can be approximated via

$$\begin{aligned} ARL_{t|30} &= \sum_{i=1}^{\infty} i\ \alpha_{t+i} \prod_{j=1}^{i-1}(1-\alpha_{t+j}) \\ &\approx \sum_{i=1}^{\infty} i\ \alpha_{30} \prod_{j=1}^{i-1}(1-\alpha_{30}) \qquad \text{for } n \geq 30 \\ &\approx 1/\alpha_{30} \qquad \text{for } n \geq 30 \end{aligned} \tag{107}$$

where α_{30} is the false alarm rate after 30 observations within the control limits. The validity of this approximation is shown in section 3.6.1. This convergence is important since it allows us to approximate the average run length for an ARMA(1,1) process by applying the recursive relationships only 30 times.

3.5.2 Initial Probability Density Functions. The second problem arises since, for the two dependent cases, the recursive relationships cannot be applied without knowing the unconditional joint probability density function, f_0. We can exploit the properties of the ARMA(1,1) process in order to obtain the unconditional joint probability density function. The ARMA(1,1) process is a special case of the general linear process. Therefore, if the errors are normally distributed, Montgomery (1991) states that arbitrary observations from an ARMA(1,1) process will also be distributed according to a normal distribution. The variance of the distribution is given as (Box and Jenkins, 1976)

$$\sigma_x^2 = \sigma_\epsilon^2 \frac{1 + \theta^2 + 2\phi\theta}{1 - \phi^2} \tag{108}$$

where σ_ϵ^2 is the variance of the independent normally distributed error term and ϕ and θ are the parameters of the ARMA(1,1) model. Then, assuming that the errors are normally distributed, the unconditional marginal probability density function of an arbitrary observation, denoted $f_0^*(x)$, is

$$f_0^*(x) = \frac{1}{\sqrt{2\pi}\sigma_\epsilon} \mathrm{e}^{-\frac{1}{2}(\frac{x}{\sigma_x})^2}. \tag{109}$$

The marginal distribution is sufficient to apply the recursive relationship for the AR(1) case. However, for the full ARMA(1,1) case, we will need the unconditional joint probability density function. The unconditional joint probability density function can be found by noting that the marginal distribution of the error is also normally distributed, implying that the joint distribution is bivariate normal

$$f_0(x, \epsilon) = \frac{1}{2\pi\sqrt{1-\rho^2}\sigma_x\sigma_\epsilon} exp[\frac{-1}{2(1-\rho^2)}(\frac{x^2}{\sigma_x^2} - \frac{2\rho x\epsilon}{\sigma_x\sigma_\epsilon} + \frac{\epsilon^2}{\sigma_\epsilon^2})] \tag{110}$$

where

$$\begin{aligned}
\rho &= Cov(X_t, \mathcal{E}_t)/\sigma_x\sigma_\epsilon \\
&= E[(X_t - \mu_x)(\mathcal{E}_t - \mu_\epsilon)]/\sigma_x\sigma_\epsilon \\
&= E[X_t\mathcal{E}_t]/\sigma_x\sigma_\epsilon \\
&= E[(\xi + \phi X_{t-1} - \theta\mathcal{E}_{t-1} + \mathcal{E}_t)\mathcal{E}_t]/\sigma_x\sigma_\epsilon \\
&= E[\mathcal{E}_t^2]/\sigma_x\sigma_\epsilon \\
&= \sigma_\epsilon^2/\sigma_x\sigma_\epsilon \\
&= \sigma_\epsilon/\sigma_x \qquad (111)
\end{aligned}$$

.

3.5.3 Solving the Recursive Relationships. The final problem arises since the recursive equations cannot be directly solved. A variety of numerical methods can be used to solve those equations given the initial probability density functions. The Matlab code used to generate the results documented in this paper for the mixed ARMA(1,1) model (case 2) is contained in Appendix B and for the pure AR(1) model (case 3) in Appendix C.

3.6 Results for False Alarm Rates.

In this section, we present the effects of autocorrelation on the false alarm rates of the X-chart. As discussed in Chapter II, quality practitioners construct control charts based, in part, on a false alarm rate which is deemed acceptable for their process. An economic tradeoff is made between the false alarm rate and the power of the control chart to detect the introduction of an assignable cause.

The traditional $\pm 3\sigma_x$ control limits for an X-chart on independent normally distributed observations result in a false alarm rate of 1/370.4 or .0027 and an average run length of

370.4. Since σ_x is generally not known, the X-chart is constructed using an estimate of the standard deviation of the error term, σ_ϵ. For the independent case, σ_ϵ is equal to the standard deviation of the measurements, σ_x.

However, when process measurements are autocorrelated, σ_ϵ is not necessarily equal to σ_x. For a stationary ARMA(1,1) model, $\sigma_\epsilon < \sigma_x$ (see equation 108). Unconditional observations from an ARMA(1,1) process with normally distributed error terms are normally distributed with a variance of σ_x^2. These observations will fall outside of control limits set at $\pm 3\sigma_x$ with a probability of 1/370.4. On the other hand, control limits set at $\pm 3\sigma_\epsilon$ will necessarily be narrower and, therefore, these observations will have a higher false alarm rate. With the exception of the independent and identically distributed case, it turns out that neither $\pm 3\sigma_x$ nor $\pm 3\sigma_\epsilon$ control limits achieve a false alarm rate of 1/370.4.

The conditional false alarm rate for an autocorrelated process with fixed control limits changes as additional conditional information about the process is acquired. The basic unit of conditional information used in this chapter is whether or not an observation is within the control limits. Since a process should be stopped to search for an assignable cause of variation whenever an observation plots outside of the control limits, a practical summary of the conditional information is simply the number of consecutive observations within the specified control limits since the process was started. **Initial control length** is defined as the number of consecutive observations that have fallen within the specified control limits to date.

3.6.1 Tables of False Alarm Rates. The change that occurs in the theoretical conditional false alarm rate is illustrated in Table 2. This table shows the theoretical false alarm rate and inverse false alarm rate for three ARMA(1,1) models with various initial control lengths. The control limits were specified at $\pm 3\sigma_x$. These limits were chosen to give equal false alarm rates when no conditional information is known. That is, the false alarm rates with no initial control points are equal for all three models as can be seen in the first row of the table. Numerical approximations of $g_i^*(x)$ and $g_i(x, \epsilon)$ were used to generate the theoretical false rates in the table. The theoretical false alarm rates were generated by applying equation 100 after each iteration of equation 97.

Similar information can be derived by simulation. Each simulated run begins with a random draw from the unconditional distribution of the ARMA model. As long as the successive observations remain within the control limits, the ARMA model is applied to extend the time series. A maximum of 30 additional observations are generated for each run. The false alarm rate for an initial control length, say n, is simply given as the number of runs in which the first n observations are within the control limits but the $n + 1$th observation is outside of the control limits divided by the number of runs in which the first n observations are within control limits. The results generated from one million simulated runs are listed in Table 3 on page 72.

The benefit provided by the theoretical approach can be appreciated by comparing the results in Tables 2 and 3 (on pages 71 and 72). This information is graphically depicted in Figure 9 on page 70. Both the theoretical and simulated results portray the same basic pattern. However, the simulated results are afflicted with additional 'noise'. Since the theoretical results are actually a numerical approximation, another important benefit of comparing the theoretical results to simulated results is to verify the accuracy of the theoretical results.

3.7 Results for Average Run Lengths.

In the previous section we saw that the false alarm rate may change dramatically prior to approaching its limiting value. Figure 9 provides strong visual evidence of the changes that may occur in the false alarm rate. In the absence of any changes to the process, the false alarm rate remains approximately constant after nearing its limiting value. Thus, the average run length from that point can be calculated as the inverse of the limiting false

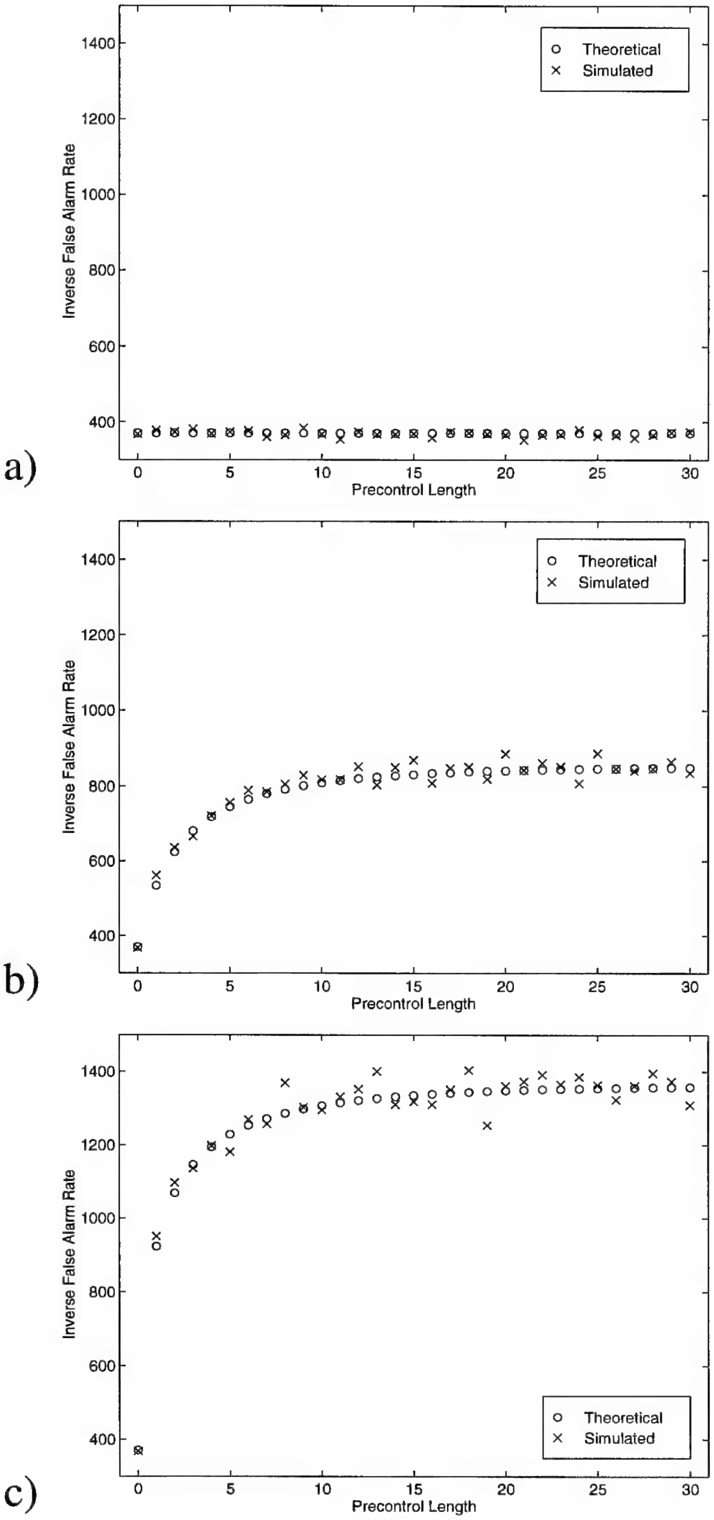


Figure 9. Inverse false alarm rates for various initial control lengths with $3\sigma_x$ control limits. ARMA(1,1) with a) $\phi = .0$ and $\theta = .0$ b) $\phi = .95$ and $\theta = .45$ c) $\phi = .95$ and $\theta = .0$.

Table 2. Theoretical false alarm rates and inverse false alarm rates for various ARMA models and various initial control lengths.

Initial Control Length	Independent Normal $\phi = 0$, $\theta = 0$		Mixed ARMA(1,1) $\phi = 0.95$, $\theta = 0.45$		AR(1) $\phi = 0.95$, $\theta = 0$	
n	α_n	$(1/\alpha_n)$	α_n	$(1/\alpha_n)$	α_n	$(1/\alpha_n)$
0	0.00270	370.4	0.00270	370.5	0.00270	370.4
1	0.00270	370.4	0.00187	533.9	0.00108	923.3
2	0.00270	370.4	0.00160	624.4	0.00094	1069.0
3	0.00270	370.4	0.00147	680.9	0.00087	1146.4
4	0.00270	370.4	0.00139	718.8	0.00084	1195.1
5	0.00270	370.4	0.00134	745.6	0.00081	1228.7
6	0.00270	370.4	0.00131	765.4	0.00080	1253.2
7	0.00270	370.4	0.00128	780.4	0.00079	1271.8
8	0.00270	370.4	0.00126	792.1	0.00078	1286.2
9	0.00270	370.4	0.00125	801.4	0.00077	1297.8
10	0.00270	370.4	0.00124	809.0	0.00077	1307.1
11	0.00270	370.4	0.00123	815.2	0.00076	1314.8
12	0.00270	370.4	0.00122	820.4	0.00076	1321.3
13	0.00270	370.4	0.00121	824.7	0.00075	1326.7
14	0.00270	370.4	0.00121	828.3	0.00075	1331.2
15	0.00270	370.4	0.00120	831.4	0.00075	1335.1
16	0.00270	370.4	0.00120	834.1	0.00075	1338.5
17	0.00270	370.4	0.00120	836.4	0.00075	1341.4
18	0.00270	370.4	0.00119	838.3	0.00074	1343.9
19	0.00270	370.4	0.00119	840.0	0.00074	1346.1
20	0.00270	370.4	0.00119	841.5	0.00074	1348.0
21	0.00270	370.4	0.00119	842.8	0.00074	1349.6
22	0.00270	370.4	0.00118	843.9	0.00074	1351.1
23	0.00270	370.4	0.00118	844.9	0.00074	1352.3
24	0.00270	370.4	0.00118	845.7	0.00074	1353.4
25	0.00270	370.4	0.00118	846.5	0.00074	1354.4
26	0.00270	370.4	0.00118	847.1	0.00074	1355.3
27	0.00270	370.4	0.00118	847.7	0.00074	1356.0
28	0.00270	370.4	0.00118	848.2	0.00074	1356.7
29	0.00270	370.4	0.00118	848.6	0.00074	1357.3
30	0.00270	370.4	0.00118	849.0	0.00074	1357.8

Table 3. Simulated false alarm rates and inverse false alarm rates for various ARMA models and various initial control lengths.

Initial Control Length	Independent Normal $\phi = 0$, $\theta = 0$		Mixed ARMA(1,1) $\phi = 0.95$, $\theta = 0.45$		AR(1) $\phi = 0.95$, $\theta = 0$	
n	α_n	$(1/\alpha_n)$	α_n	$(1/\alpha_n)$	α_n	$(1/\alpha_n)$
0	0.00272	368.1	0.00272	368.1	0.00272	368.1
1	0.00264	378.2	0.00178	561.5	0.00105	950.7
2	0.00268	373.6	0.00157	635.7	0.00091	1097.2
3	0.00263	380.9	0.00150	666.2	0.00088	1136.2
4	0.00271	369.2	0.00139	721.8	0.00083	1198.1
5	0.00268	373.5	0.00132	756.5	0.00085	1181.5
6	0.00266	376.5	0.00127	789.3	0.00079	1269.5
7	0.00277	360.4	0.00127	785.2	0.00080	1257.3
8	0.00273	365.7	0.00124	805.3	0.00073	1369.1
9	0.00260	384.0	0.00121	829.3	0.00077	1303.3
10	0.00272	368.2	0.00122	817.3	0.00077	1295.5
11	0.00282	354.1	0.00122	817.6	0.00075	1331.0
12	0.00267	373.9	0.00117	852.1	0.00074	1351.9
13	0.00272	367.7	0.00124	803.7	0.00071	1400.7
14	0.00272	367.8	0.00118	850.0	0.00076	1310.5
15	0.00271	369.2	0.00115	869.4	0.00076	1318.2
16	0.00279	358.5	0.00124	808.7	0.00076	1310.2
17	0.00268	373.6	0.00118	847.7	0.00074	1350.5
18	0.00269	371.2	0.00117	851.1	0.00071	1403.4
19	0.00271	368.7	0.00122	819.4	0.00080	1254.0
20	0.00273	366.7	0.00113	885.4	0.00074	1360.5
21	0.00284	352.1	0.00119	843.0	0.00073	1372.9
22	0.00273	366.0	0.00116	861.4	0.00072	1391.3
23	0.00272	367.1	0.00117	852.1	0.00073	1365.1
24	0.00264	379.3	0.00124	807.8	0.00072	1385.4
25	0.00275	363.0	0.00113	886.7	0.00073	1363.2
26	0.00274	365.4	0.00118	846.9	0.00076	1323.4
27	0.00280	357.7	0.00119	840.7	0.00073	1361.1
28	0.00273	366.0	0.00118	846.3	0.00072	1395.1
29	0.00268	372.5	0.00116	864.3	0.00073	1372.5
30	0.00267	374.2	0.00120	835.6	0.00076	1308.9

alarm rate via equation 107. In this section, we present additional results that further verify these conjectures.

Using the same techniques as in the previous section, the probability density function for the next observation after a initial control length of 30 observations can be numerically approximated. The theoretical false alarm rate can then be computed by integrating that probability density function outside of the control limits. A theoretical average run length can then be computed as the reciprocal of the false alarm rate. Table 4 on page 75 contains the theoretical average run length for a variety of models and a variety of control limit multipliers. For ease of comparison to other results in the literature, this table mirrors the layout used in Wardell, Moskowitz and Plante (1994) .

As in the previous section, we can both verify the numerically approximated theoretical results and demonstrate their value by generating comparable results using simulation. Table 5 on page 76 mirrors Table 4 and contains the mean run length from 10,000 simulated runs for for each combination of ARMA parameters and control limits. As in the previous section, each run began with a random draw from the unconditional distribution of the ARMA(1,1) process. Additional observations were generated to achieve a initial control length of 30. If any observation fell outside of the control limits in this phase, the run was restarted. Once a initial control length of 30 observations was achieved, the time series was extended until an observation was generated outside of the control limits. The value listed in the table is the arithmetic mean of the run lengths achieved. The limits for a (1-α)100 percent confidence interval can be constructed around each simulated average run length via

$$ARL \pm t_{\alpha/2}\frac{SRL}{\sqrt{n}} \tag{112}$$

where $t_{\alpha/2}$ is the appropriate value of the t distribution with $n-1$ degrees of freedom, ARL is the average run length, and SRL is the sample standard deviation of the simulated run lengths. Tables 6 and 7 (on pages 77 and 78) contain lower and upper limits for a 90 percent confidence interval about the average run length. The numerically approximated theoretical average run length is within the confidence interval for 175 out of the 200 design points, or, for 87.5 percent of the design points. Furthermore, the theoretical average run length is within a 99 percent confidence interval for every design point. These results confirm that the numerically approximated theoretical average run lengths are not inconsistent with the average run lengths from a large scale simulation effort. That is, the statistical evidence tends to confirm that the numerically approximated average run lengths are correct.

While the results in Tables 4 or 5 allow the average run length for a variety of ARMA(1,1) processes to be identified, it would be more practical for somebody setting up a control chart to have a table indexed on the average run length. Three average run lengths are used in Table 8: 110, 370 and 1000. An average run length of 370 is commonly used due to the prevalence of $\pm 3\sigma$ control limits on independent normal data. Some authors suggest using the more responsive average run length of 110 (Wardell et al., 1992). We also include an average run length of 1000 corresponding to a false alarm rate of .01 percent for independent observations. The control limit multipliers presented in Table 8 were found by interpolating the results in Table 4 on a log-log scale. The use of Table 8 is discussed in the next section.

3.8 Using the Results for Quality Control.

A quality control practitioner with an autocorrelated process may be unwilling to use standard control charts due to the unacceptable increase in the false alarm rate. The research in this chapter provides a basis for selecting control limits for a known ARMA(1,1) process to achieve a desired average run length in the absence of any assignable causes of variation. Although it can be argued that we never have a known ARMA(1,1) process, we typically assume that our process can be adequately modelled as a ARMA(1,1) process.

Suppose we want to select control limits for a process that can be approximated by an ARMA(1,1) process with mean μ and parameters ϕ and θ. The control limits corresponding to a variety of average run lengths can be constructed by selecting the appropriate control

Table 4. Numerically Approximated Average Run Length for control limits set at various multiples of σ_x after 30 initial control observations.

	L_x	ARL_{30}				
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.90	1.50	7.6	8.0	8.9	11.6	50.2
	1.75	12.8	13.2	14.7	18.9	82.6
	2.00	22.4	23.0	25.1	32.3	138.5
	2.25	41.6	42.3	46.1	57.9	239.4
	2.50	81.4	82.3	87.7	109.3	430.2
	2.75	168.7	169.9	180.7	218.1	812.3
	3.00	370.7	372.4	386.6	464.3	1622.9
	3.25	863.6	866.6	901.7	1043.6	3440.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.45	1.50	26.7	7.5	8.2	10.8	48.8
	1.75	44.9	12.5	13.7	17.7	80.4
	2.00	75.8	22.0	23.7	30.4	134.7
	2.25	130.6	40.9	43.4	54.5	232.4
	2.50	232.4	80.5	84.7	103.2	417.3
	2.75	432.3	167.6	173.6	206.9	786.4
	3.00	849.0	370.1	377.5	440.1	1567.3
	3.25	1772.1	866.4	883.8	996.5	3318.9
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.0	1.50	43.0	8.9	7.5	8.9	43.0
	1.75	71.0	14.7	12.5	14.7	71.0
	2.00	118.8	25.4	22.0	25.4	118.8
	2.25	204.5	46.3	40.9	46.3	204.5
	2.50	365.6	89.0	80.5	89.0	365.6
	2.75	685.6	181.3	167.8	181.3	685.6
	3.00	1358.9	392.4	370.4	392.4	1358.9
	3.25	2862.7	903.3	866.5	903.3	2862.7
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ -0.45	1.50	48.8	10.8	8.2	7.5	26.7
	1.75	80.4	17.7	13.7	12.5	44.9
	2.00	134.7	30.4	23.7	22.0	75.8
	2.25	232.4	54.5	43.4	40.9	130.6
	2.50	417.3	103.2	84.7	80.5	232.4
	2.75	786.4	206.9	173.6	167.6	432.3
	3.00	1567.3	440.1	377.5	370.1	849.0
	3.25	3318.9	996.5	883.8	866.4	1772.1
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ -0.90	1.50	50.2	11.6	8.9	8.0	7.6
	1.75	82.6	18.9	14.7	13.2	12.8
	2.00	138.5	32.3	25.1	23.0	22.4
	2.25	239.4	57.9	46.1	42.3	41.6
	2.50	430.2	109.3	87.7	82.3	81.4
	2.75	812.3	218.1	180.7	169.9	168.7
	3.00	1622.9	464.3	386.6	372.4	370.7
	3.25	3440.6	1043.6	901.7	866.6	863.6

Table 5. Average Run Length from simulation for control limits set at various multiples of σ_x after 30 initial control observations.

	L_x	ARL_{30}				
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.90	1.50	7.7	7.9	8.9	11.7	49.7
	1.75	12.8	13.3	14.6	18.8	83.0
	2.00	22.9	23.2	25.8	32.8	139.4
	2.25	41.5	42.3	46.4	58.0	235.8
	2.50	80.6	83.7	88.3	109.7	431.2
	2.75	169.8	172.9	177.5	217.4	810.3
	3.00	376.8	374.3	394.5	462.1	1623.1
	3.25	879.9	863.4	877.0	1035.4	3531.9
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.45	1.50	26.8	7.4	8.2	10.8	48.1
	1.75	44.5	12.6	13.7	17.9	80.1
	2.00	76.5	22.2	23.6	30.2	135.4
	2.25	129.8	41.1	43.4	54.8	231.8
	2.50	231.5	80.5	84.1	103.1	419.2
	2.75	436.3	166.2	177.4	205.0	779.2
	3.00	867.5	377.1	374.3	441.1	1555.1
	3.25	1791.2	859.3	874.1	986.8	3382.3
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.00	1.50	43.1	8.9	7.5	8.8	42.4
	1.75	69.8	14.6	12.5	15.0	69.7
	2.00	119.8	25.4	22.3	25.6	119.0
	2.25	202.2	46.1	41.3	46.6	202.1
	2.50	363.5	90.0	80.1	88.5	363.4
	2.75	687.3	182.6	166.0	180.7	675.9
	3.00	1353.3	398.4	374.7	392.7	1358.1
	3.25	2909.7	901.4	860.8	905.1	2841.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ -0.45	1.50	49.2	11.0	8.4	7.4	26.3
	1.75	79.4	17.8	14.0	12.6	44.4
	2.00	135.7	30.4	24.0	22.2	75.4
	2.25	231.0	54.8	42.8	41.4	131.9
	2.50	421.1	103.8	87.1	79.9	230.9
	2.75	789.5	209.5	172.7	165.1	439.1
	3.00	1588.5	450.4	372.5	374.3	843.8
	3.25	3409.6	995.6	897.0	865.5	1783.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.90	1.50	50.7	11.8	9.1	8.1	7.7
	1.75	81.9	18.9	14.9	13.1	12.9
	2.00	138.0	31.5	25.7	23.2	22.7
	2.25	239.2	57.3	45.9	42.9	41.7
	2.50	433.8	109.3	88.2	81.7	81.5
	2.75	814.8	219.5	182.1	170.7	166.0
	3.00	1635.3	473.7	394.8	374.3	380.4
	3.25	3540.3	1041.3	915.5	871.0	876.0

Table 6. Lower limit of a 90 percent confidence interval for the Average Run Length from simulation for control limits set at various multiples of σ_x after 30 initial control observations.

	L_x	ARL_{30}				
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	1.50	7.6	7.8	8.8	11.5	48.9
	1.75	12.6	13.1	14.3	18.5	81.6
$\theta=$	2.00	22.5	22.8	25.4	32.3	137.1
0.90	2.25	40.9	41.6	45.6	57.1	231.9
	2.50	79.3	82.3	86.9	107.9	424.2
	2.75	166.9	170.1	174.6	213.9	797.1
	3.00	370.7	368.1	388.0	454.4	1596.8
	3.25	865.4	849.3	862.5	1018.6	3474.7
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	1.50	26.4	7.3	8.1	10.6	47.3
	1.75	43.8	12.4	13.4	17.6	78.8
$\theta=$	2.00	75.2	21.8	23.3	29.7	133.2
0.45	2.25	127.7	40.5	42.7	53.9	228.0
	2.50	227.7	79.2	82.8	101.5	412.4
	2.75	429.1	163.6	174.5	201.7	766.5
	3.00	853.1	371.0	368.2	433.9	1529.7
	3.25	1761.9	845.1	860.0	970.8	3327.1
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	1.50	42.3	8.8	7.4	8.7	41.8
	1.75	68.6	14.4	12.3	14.7	68.6
$\theta=$	2.00	117.9	25.0	22.0	25.2	117.1
0.00	2.25	198.8	45.3	40.7	45.9	198.7
	2.50	357.5	88.5	78.8	87.1	357.5
	2.75	675.9	179.6	163.3	177.7	665.0
	3.00	1331.2	391.8	368.7	386.3	1336.0
	3.25	2861.9	886.7	846.5	890.6	2794.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	1.50	48.4	10.8	8.3	7.3	25.9
	1.75	78.1	17.6	13.8	12.4	43.7
$\theta=$	2.00	133.5	29.9	23.6	21.8	74.2
-0.45	2.25	227.2	53.9	42.1	40.8	129.8
	2.50	414.2	102.1	85.7	78.6	227.1
	2.75	776.6	206.1	169.9	162.5	431.9
	3.00	1562.7	443.0	366.2	368.2	830.0
	3.25	3352.9	978.9	882.4	851.1	1754.5
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	1.50	49.9	11.6	8.9	7.9	7.6
	1.75	80.6	18.6	14.7	12.9	12.7
$\theta=$	2.00	135.7	31.0	25.3	22.8	22.3
0.90	2.25	235.3	56.4	45.2	42.2	41.1
	2.50	426.7	107.5	86.8	80.3	80.2
	2.75	801.4	215.9	179.1	167.8	163.3
	3.00	1608.6	466.1	388.2	368.0	374.2
	3.25	3481.0	1023.9	900.4	856.4	861.6

Table 7. Upper limit of a 90 percent confidence interval for the Average Run Length from simulation for control limits set at various multiples of σ_x after 30 initial control observations.

	L_x	ARL_{30}				
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.90	1.50	7.8	8.0	9.1	11.9	50.5
	1.75	13.0	13.5	14.8	19.1	84.3
	2.00	23.3	23.5	26.3	33.4	141.6
	2.25	42.2	43.0	47.1	59.0	239.7
	2.50	81.9	85.0	89.8	111.5	438.2
	2.75	172.6	175.8	180.5	220.9	823.5
	3.00	382.8	380.4	401.0	469.7	1649.5
	3.25	894.3	877.5	891.5	1052.2	3589.0
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.45	1.50	27.2	7.6	8.3	11.0	48.8
	1.75	45.2	12.8	13.9	18.2	81.4
	2.00	77.7	22.5	24.0	30.7	137.6
	2.25	131.9	41.8	44.2	55.6	235.6
	2.50	235.3	81.8	85.5	104.8	426.0
	2.75	443.6	168.9	180.3	208.4	791.8
	3.00	881.9	383.2	380.5	448.2	1580.4
	3.25	1820.5	873.4	888.2	1002.8	3437.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.00	1.50	43.8	9.1	7.6	8.9	43.1
	1.75	70.9	14.8	12.7	15.2	70.9
	2.00	121.8	25.8	22.7	26.0	121.0
	2.25	205.6	46.8	42.0	47.4	205.4
	2.50	369.5	91.5	81.4	90.0	369.3
	2.75	698.7	185.5	168.7	183.7	686.8
	3.00	1375.3	405.1	380.8	399.2	1380.3
	3.25	2957.4	916.1	875.1	919.6	2888.6
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ -0.45	1.50	50.0	11.2	8.6	7.5	26.8
	1.75	80.7	18.1	14.2	12.8	45.2
	2.00	137.8	30.9	24.4	22.6	76.7
	2.25	234.8	55.7	43.5	42.1	134.1
	2.50	428.1	105.5	88.5	81.2	234.6
	2.75	802.5	213.0	175.5	167.8	446.2
	3.00	1614.4	457.7	378.8	380.3	857.6
	3.25	3466.3	1012.3	911.5	879.9	1812.7
		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta=$ 0.90	1.50	51.6	12.0	9.2	8.2	7.8
	1.75	83.2	19.2	15.2	13.4	13.1
	2.00	140.2	32.0	26.1	23.5	23.0
	2.25	243.2	58.2	46.7	43.6	42.4
	2.50	440.9	111.1	89.6	83.0	82.8
	2.75	828.3	223.2	185.1	173.5	168.7
	3.00	1662.1	481.3	401.4	380.6	386.6
	3.25	3599.6	1058.7	930.6	885.5	890.4

Table 8. Equivalent control limit multipliers of σ_x and σ_ϵ corresponding to desired average run lengths after 30 initial control observations.

		$\phi = 0.95$		$\phi = 0.475$		$\phi = 0.0$		$\phi = -0.475$		$\phi = -0.95$	
	Desired ARL	L_x	L_ϵ	L_x	L_ϵ	L_x	L_ϵ	L_x	L_ϵ	L_x	L_ϵ
$\theta =$ 0.90	110	2.61	2.64	2.60	2.89	2.58	3.47	2.50	4.64	1.89	11.36
	370	3.00	3.04	3.00	3.33	2.99	4.02	2.93	5.43	2.44	14.65
	1000	3.29	3.33	3.29	3.65	3.28	4.41	3.24	6.01	2.83	16.99
$\theta =$ 0.45	110	2.17	4.10	2.61	2.61	2.59	2.84	2.52	3.66	1.90	8.74
	370	2.69	5.08	3.00	3.00	2.99	3.28	2.94	4.27	2.45	11.26
	1000	3.06	5.77	3.29	3.29	3.28	3.60	3.25	4.72	2.84	13.04
$\theta =$ 0.00	110	1.96	6.29	2.58	2.93	2.61	2.61	2.58	2.93	1.96	6.29
	370	2.51	8.02	2.98	3.39	3.00	3.00	2.98	3.39	2.51	8.02
	1000	2.89	9.26	3.28	3.73	3.29	3.29	3.28	3.73	2.89	9.26
$\theta =$ -0.45	110	1.90	8.74	2.52	3.66	2.59	2.84	2.61	2.61	2.17	4.10
	370	2.45	11.26	2.94	4.27	2.99	3.28	3.00	3.00	2.69	5.08
	1000	2.84	13.04	3.25	4.72	3.28	3.60	3.29	3.29	3.06	5.77
$\theta =$ -0.90	110	1.89	11.36	2.50	4.64	2.58	3.47	2.60	2.89	2.61	2.64
	370	2.44	14.65	2.93	5.43	2.99	4.02	3.00	3.33	3.00	3.04
	1000	2.83	16.99	3.24	6.01	3.28	4.41	3.29	3.65	3.29	3.33

limit multiplier from Table 8. The table includes three generally accepted average run lengths: 110, 370 and 1000. The control limits can be calculated via

$$\begin{aligned} UCL &= \mu + L_x\sigma_x \\ LCL &= \mu - L_x\sigma_x \end{aligned} \tag{113}$$

or

$$\begin{aligned} UCL &= \mu + L_\epsilon\sigma_\epsilon \\ LCL &= \mu - L_\epsilon\sigma_\epsilon. \end{aligned} \tag{114}$$

The tables implicitly assumes a initial control length of at least 30. This requirement is reasonable in practice since the ARMA model parameters cannot generally be adequately estimated with less than 30 observations. If the initial control period is less than 30, the average run length will tend to be shorter than desired. On the other hand, if the process is centered prior to monitoring, the average run length should approximate the desired average run length. Of course, selecting a multiplier corresponding to a lower average run length (implying narrower control limits) can be expected to detect the introduction of an assignable cause more quickly than a multiplier corresponding to a higher average run length.

In our experience, a control limit multiplier for an ARMA(1,1) process with autoregressive and moving average parameters that are within the ranges contained in Table 8 can be succesfully determined using a cubic interpolation of the tabulated values. Of course, we do not advise extrapolating outside of the range of parameters listed in the table. Nor are enough data points provided to determine control limit multipliers corresponding to desired average run lengths other than the three listed.

3.9 Verifying the Results for Quality Control

We conducted simulations using the tabled control limit multipliers to verify Table 8. An estimate of the average run length was computed as the arithmetic mean of the run lengths from 10,000 runs. The results of the simulations are listed in Tables 9, 10 and 11. The desired average run length was within the confidence interval in 62 out of the 75 design points, or 82.67 percent of the time. The desired average run length was outside of a 99 percent confidence interval at only one design point (for the desired average run length of 1000 with ϕ = -0.475 and θ = 0.45). The simulated average run length for this case was 1027.8 and had a 99 percent confidence interval of (1001.3, 1054.3). These observations demonstrate that there is no strong statistical evidence to say that the tabulated control limit multipliers do not generate the corresponding average run lengths. That is, the results of the simulated runs tend to confirm that the tabulated control limit multipliers do generate the corresponding average run lengths.

Table 9. Simulation results for control limit multipliers of σ_x corresponding to desired average run lengths after 30 initial control observations.

		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	Desired ARL	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$
$\theta=$ 0.90	110	112.1	112.4	111.5	111.0	110.9
	370	364.8	372.1	376.4	369.6	371.4
	1000	1023.4	1019.7	995.5	1004.3	995.5
$\theta=$ 0.45	110	110.5	112.0	110.5	112.0	109.8
	370	371.8	364.3	370.1	369.4	371.2
	1000	1008.4	1007.4	1003.5	1027.8	994.3
$\theta=$ 0.00	110	109.4	110.2	111.7	110.7	109.1
	370	368.5	367.6	364.6	367.4	370.1
	1000	1012.9	996.1	1010.0	1016.4	994.6
$\theta=$ -0.45	110	110.4	112.1	110.1	111.3	110.9
	370	371.0	368.9	373.6	364.3	375.0
	1000	1015.1	1025.7	990.0	1007.3	1002.7
$\theta=$ 0.90	110	111.0	111.7	110.9	111.4	111.5
	370	369.0	367.1	377.4	368.6	372.0
	1000	1019.6	1024.8	999.7	1005.6	991.8

3.10 Chapter Summary.

Using standard techniques to select fixed control limits for an autocorrelated process results in unpredictably high false alarm rates. The false alarm rates are manifested by a shortened average run length in the absence of assignable causes of variation and, in turn, an excessive number of searches for nonexistent assignable causes. In this chapter, we developed and tested a method for selecting fixed control limits for an ARMA(1,1) model to achieve a specified average run length in the absence of assignable cause variation. The method is an improvement over existing techniques since it provides a known average run length and should, therefore, reduce the number of searches for nonexistent assignable causes. We used the method to develop a table of control limit multipliers corresponding to a diverse set of ARMA(1,1) models that achieve standard average run lengths. A quality practitioner with a process, approximated by an ARMA(1,1) model, can identify control limits corresponding to a given average run length by interpolating the tabulated results.

Table 10. Lower limit of 90 percent confidence interval on average run length after 30 initial control observations.

		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	Desired ARL	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$
$\theta=$ 0.90	110	110.2	110.6	109.7	109.1	109.2
	370	358.9	366.1	370.1	363.6	365.4
	1000	1006.5	1003.4	979.6	987.8	979.2
$\theta=$ 0.45	110	108.7	110.2	108.7	110.1	108.0
	370	365.7	358.3	364.0	363.2	365.1
	1000	991.8	991.0	987.4	1011.0	978.1
$\theta=$ 0.00	110	107.6	108.4	109.8	108.9	107.3
	370	362.5	361.7	358.6	361.4	364.1
	1000	996.3	979.9	993.6	999.7	978.4
$\theta=$ -0.45	110	108.5	110.3	108.3	109.5	109.1
	370	364.9	362.9	367.6	358.3	368.8
	1000	998.3	1008.8	973.7	990.8	986.3
$\theta=$ 0.90	110	109.2	109.9	109.1	109.6	109.6
	370	362.9	361.1	371.2	362.5	365.9
	1000	1002.7	1007.9	983.4	989.0	975.8

Table 11. Upper limit of 90 percent confidence interval on average run length after 30 initial control observations.

		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
	Desired ARL	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$	$A\hat{R}L$
$\theta=$ 0.90	110	113.9	114.2	113.3	112.8	112.7
	370	370.8	378.2	382.6	375.7	377.5
	1000	1040.3	1035.9	1011.5	1020.7	1011.8
$\theta=$ 0.45	110	112.3	113.9	112.3	113.8	111.6
	370	377.9	370.3	376.2	375.5	377.2
	1000	1024.9	1023.8	1019.6	1044.7	1010.5
$\theta=$ 0.00	110	111.1	112.0	113.5	112.6	110.9
	370	374.6	373.6	370.6	373.4	376.2
	1000	1029.5	1012.4	1026.5	1033.0	1010.8
$\theta=$ -0.45	110	112.2	114.0	111.9	113.2	112.8
	370	377.1	374.8	379.7	370.3	381.1
	1000	1032.0	1042.6	1006.3	1023.7	1019.2
$\theta=$ 0.90	110	112.9	113.6	112.7	113.2	113.3
	370	375.1	373.0	383.6	374.7	378.1
	1000	1036.5	1041.8	1016.0	1022.1	1007.8

IV. A Theoretical Foundation for Assessing Process Capability

The majority of statistical process control techniques in use today are based upon monitoring the state of statistical control. We presented the tools and techniques currently used to do so in the first two chapters of this dissertation. In the previous chapter, we presented an extension to the Shewhart control chart to account for autocorrelated process observations. Beginning in this chapter, the focus of our research shifts to a new paradigm based upon monitoring the capability of a process over time rather than monitoring its state of statistical control.

The conceptual shift to a new paradigm requires a theoretical foundation to build upon. We develop that foundation in this chapter. The theory we develop in this chapter departs from the existing literature by emphasizing the notion that the capability of a process can vary over time. The time-varying aspect of capability can be used as a means for process control. Furthermore, in this chapter we show that the capability of a process at some time in the future can be estimated by fitting a model to a known set of observations from that process. In the next chapter, we apply these theoretical concepts in the development of a practical method for monitoring process capability.

4.1 Overview

In the first part of this chapter, we extend the definition of capability to account for the long-term aspect of capability implied in the existing definitions as well as the time-varying aspects of capability developed in this chapter. We illustrate these aspects of capability by applying them to three cases: an independent and identically distributed process; a deterministic sinusoidal process; and a more general process. Next, we develop a method for estimating the process capability for the very next observation for each of the three cases. The results derived for estimating capability one time step into the future are followed by more general results for estimating process capability at an arbitrary number of time steps into the future.

4.2 Assumptions, Notation and Definitions

The definitions and assumptions prevalent in the statistical process control literature are centered around the paradigm of monitoring the state of statistical control. In this section, we present a minimal set of background assumptions which will be used within the rest of the chapter. We also make the distinction between time-specific, time-average and long-term capability. Finally, we define time-specific expected loss, time-average expected loss and long-term expected loss.

4.2.1 Background Assumptions and Notation. Consider a process with some measure of quality evaluated at discrete time intervals. Given the N most recent process observations at time T, the period of time during which the future behavior of the process is of interest is defined by the index set $\{T+1, T+2, \ldots, T+s\}$, where $T+s$ is the time associated with last observation of interest. The measure of quality at some time, $T+k$, in the index set can be considered to be an observation from a random variable, denoted X_{T+k}, whose (conditional) probability density function is denoted $f_{T+k|T}$. The conditional expected value of the process at time $T+k$ given the N most recent observations at time T, denoted $\mu_{T+k|T}$, is given by

$$\begin{aligned} \mu_{T+k|T} &\triangleq E[X_{T+k}|x_1, \ldots, x_N] \\ &= \int_{-\infty}^{\infty} x f_{T+k|T}(x)\, dx. \end{aligned} \tag{115}$$

Note the minor notational shorthand: $E(X_{T+k}|x_1, \ldots, x_N)$ is used instead of the more cumbersome $E(X_{T+k}|X_1 = x_1, \ldots, X_N = x_N)$. The variance of the process at time $T+k$

given the observations up to time T, denoted $\sigma^2_{T+k|T}$, is similarly given by

$$\begin{aligned}
\sigma^2_{T+k|T} &\triangleq Var[X_{T+k}|x_1, \ldots, x_N] \\
&= E[X^2_{T+k}|x_1, \ldots, x_N] - E^2[X_{T+k}|x_1, \ldots, x_N] \\
&= E[X^2_{T+k}|x_1, \ldots, x_N] - \mu^2_{T+k|T}. \qquad (116)
\end{aligned}$$

4.2.2 Time-specific, Time-average and Long-term Capability Defined. Capability, as defined in Chapter I, is a measure of the ability of a process to produce items within the process specification limits. Under this definition, one process is more capable than another if it produces fewer items outside of its specification limits. While this definition is sufficient for independent and identically distributed process measurements, it is ambiguous when applied to more complicated real world processes. These modern processes, characterized by autocorrelated measurements, may produce a high percentage of items within the specification limits over a long period of time while producing high numbers of items outside of the specification limits during short intervals of time. For example, consider the two processes depicted in Figure 10 on page 86. The bottom process is independent normal while the top process is AR(1) with $\phi = .97$. The variance of the independent normal errors for the AR(1) process was scaled so that the unconditional distributions for both processes are equal. The histograms on the left side of the figure provide visual evidence of that equality. However, on the right side of the figure, the tendency of the observations from the AR(1) process to vary together is evident. If the lower control limit were set to -1.5, the observations from the independent normal process would be below the lower specification limit 21 times during the 400 observations. Those observations are spread out over the 400 observations and, in general, occur independently. Observations from the AR(1) process are below the lower specification limit 12 times between observation 356 and 373, but never at any other time. In general, the observations outside of a specification limit will be clustered since the observations are positively autocorrelated. If the period

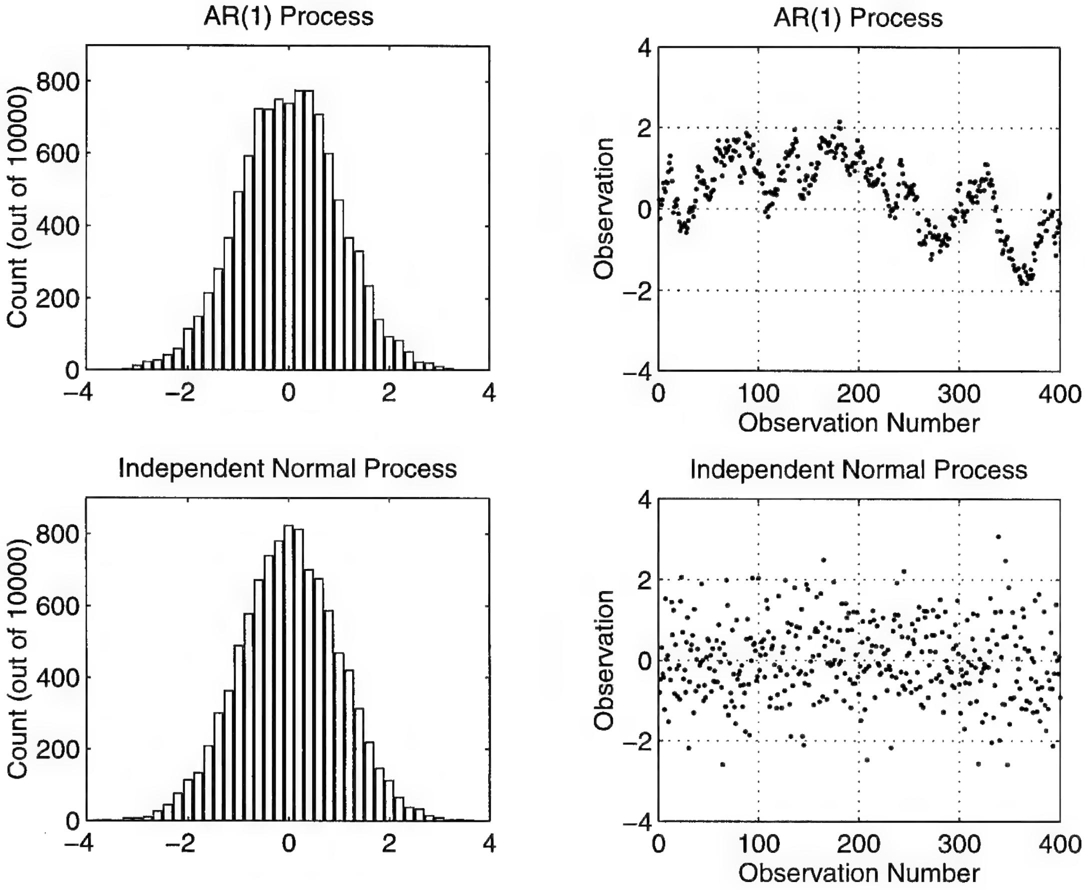


Figure 10. Comparison between an independent process and a highly autocorrelated process.

of time during which the process measurements are of interest is likely to include such a run of out-of-specification observations, the process should not be considered as capable as the independent normal process. The current definition of capability does not adequately address whether such a process is capable or not. In this subsection, we provide extended definitions that accommodate the multifaceted nature of capability.

Suppose that the quality characteristic has a specified ideal target value denoted τ, an upper specification limit denoted USL, and a lower specification limit denoted LSL. One of the key concepts proposed in this research is that capability can be viewed as varying over time. Recall from Chapter I that capability is a measure of the total variation in the process against the specifications. Under a strict interpretation of the specification limits, the capability of the process at a given time can be defined as the probability that the observation at that time will fall within the upper and lower specification limits. The *time-specific process capability* at some time, $T+k$, given the N most recent observations at time T, can be expressed as

$$\mathcal{C}_{T+k|T} \triangleq P[LSL \leq X_{T+k} \leq USL|x_1, \ldots, x_N]. \tag{117}$$

In other words, the time-specific capability at time $T+k$ is a function of the distribution of the random variable $X_{T+k|T}$, via

$$\mathcal{C}_{T+k|T} = \int_{LSL}^{USL} f_{T+k|T}(x)dx. \tag{118}$$

Although process capability may vary with time, we still want to be able to express the capability of the process over a period of time. In this case, the total variation of the output during the period of interest can be considered a combination of the variations from each time in the period of interest. Again, under a strict interpretation of the specification limits, the *time-average process capability* during the period of interest can be defined as the proportion of observations during the period of interest which are expected to fall within the upper and lower specification limits. In order to evaluate the capability of the process throughout the period of interest, we will construct a probability density function over the entire index set as a weighted linear combination of the individual probability density functions via

$$f_I(x) = 1/s \sum_{k=1}^{s} f_{T+k|T}(x). \tag{119}$$

The density function, f_I, concisely incorporates the total variation of the process output throughout the period of interest. Time-average capability is a function of the total variation throughout the period of interest, and thus, can be considered as a function of the combined density functions. Using equal weights is a natural choice, although it is not required. Unequal weights may be desired when economic concerns emphasize particular times (e.g. when added importance is placed upon the measurements in the immediate future). In this case,

$$f_I(x) = \sum_{k=1}^{s} \gamma_{T+k}\ f_{T+k|T}(x) \tag{120}$$

where

$$\sum_{k=1}^{s} \gamma_{T+k} = 1 \tag{121}$$

and $\gamma_{T+k} \geq 0$ for all $1 \leq k \leq s$. In either case, the probability density function, f_I, defines a random variable X_I. We will use the more general notation allowing for unequal weights from here on. The *time-average expected value* of the process during the period of interest, denoted μ_I, can be expressed via

$$\begin{aligned}
\mu_I &= E[X_I] \\
&= \int_{-\infty}^{\infty} x f_I(x)\ dx \\
&= \int_{-\infty}^{\infty} x \sum_{k=1}^{s} \gamma_{T+k} f_{T+k|T}(x)\ dx \\
&= \sum_{k=1}^{s} \gamma_{T+k} \int_{-\infty}^{\infty} x f_{T+k|T}(x)\ dx \\
&= \sum_{k=1}^{s} \gamma_{T+k} \mu_{T+k|T}. \qquad (122)
\end{aligned}$$

The *time-average variance* of the process during the period of interest, denoted σ_I^2, can be similarly expressed via

$$\begin{aligned}
\sigma_I^2 &= Var[X_I] \\
&= E[X_I^2] - E^2[X_I] \\
&= \int_{-\infty}^{\infty} x^2 f_I(x)\, dx - (\sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T})^2 \\
&= \int_{-\infty}^{\infty} x^2 \sum_{k=1}^{s} \gamma_{T+k} f_{T+k|T}(x)\, dx - (\sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T})^2 \\
&= \sum_{k=1}^{s} \gamma_{T+k} \int_{-\infty}^{\infty} x^2 f_{T+k|T}(x)\, dx - (\sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T})^2 \\
&= \sum_{k=1}^{s} \gamma_{T+k} E[X_{T+k|T}^2 | x_1, \ldots, x_N] - (\sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T})^2 \\
&= \sum_{k=1}^{s} \gamma_{T+k}\sigma_{T+k|T}^2 + \sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T}^2 - (\sum_{k=1}^{s} \gamma_{T+k}\mu_{T+k|T})^2. \qquad (123)
\end{aligned}$$

Using the notation developed in the previous section, the time-average process capability during the period of interest can be expressed as

$$\begin{aligned}
\mathcal{C}_I &\triangleq P[LSL \leq X_I \leq USL] \\
&= \int_{LSL}^{USL} f_I(x) dx. \qquad (124)
\end{aligned}$$

It is very important to note that $\mathcal{C}_{T+k|T}$ does not necessarily equal $\mathcal{C}_I$. Examples of this will be seen in the following sections.

When the observations throughout the index set are independent and identically distributed, $f_{T+k|T} = f_I$ for all $1 \leq k \leq s$. In this case, the conditional expected value of the process at time $T + k$ given the observations up to time T is equal to the expected value of X_I (i.e. $\mu_{T+k|T} = \mu_I$). Furthermore, the conditional variance of the process at time $T + k$ given the observations up to time T is equal to the variance of X_I (i.e. $\sigma_{T+k|T}^2 = \sigma_I^2$).

More generally, the observations over the index set are not independent and identically distributed. We do, however, typically impose a structure upon these observations. For instance, in the next chapter, process measurements are assumed to be observations from a stationary ARMA(1,1) process. Given a set of known observations and the parameters of the ARMA(1,1) model, the conditional probability density function of future measurements can be identified. In this chapter, we assume that the process is stationary, or, $f_{T+l} = f_{T+m}$ for all $T + l$ and $T + m$ in $1 \leq l \leq m \leq s$. Clearly, the conditional probability density functions of a stationary process are not necessarily equal (i.e. $f_{T+l|T} \neq f_{T+m|T}$). However, we will assume that the impact of additional information about process observations decreases over time such that, after a sufficiently large time, $f_{T+k|T}$ approaches a limiting density function f_x. Indeed, for any stationary ARMA(1,1) process, $\lim_{k\to\infty} f_{T+k|T} = f_x$, where f_x is the unconditional probability density function of the process observations. More precisely, for any x and any positive real number δ, there exists an integer K such that $|f_{T+k|T}(x) - f_x(x)| < \delta$ for every $k > K$.

When the conditional probability density functions approach a limiting function, f_x, then the function f_I also approaches f_x as the size, s, of the index set increases (i.e. $\lim_{s\to\infty} f_I = f_x$). To see this convergence, suppose we are given a value, x, in the domain of f_x and a positive real number δ. Since $f_{T+k|T}$ approaches f_x, we know there exists an integer $K1$ such that $|f_{T+k|T}(x) - f_x(x)| < \delta/2$ for all k greater than $K1$. Choose $K2$ such that

$$\sum_{k=1}^{K1} |(f_{T+k|T}(x) - f_x(x))| \leq \frac{K1 + K2}{2}\delta \qquad (125)$$

and let $N = K1 + K2$. Then, for any $s > N$, and given that $\{\gamma_{T+k}\}$ is a sequence for which $\gamma_{T+i} \leq 1/s$ for all $i < K1$,

$$\begin{aligned}
|f_I(x) - f_x(x)| &= |\sum_{k=1}^{s} \gamma_{T+k} f_{T+k|T}(x) - \sum_{k=1}^{s} \gamma_{T+k} f_x(x)| \\
&= |\sum_{k=1}^{s} \gamma_{T+k}(f_{T+k|T}(x) - f_x(x))| \\
&= |\sum_{k=1}^{K1} \gamma_{T+k}(f_{T+k|T}(x) - f_x(x)) + \sum_{k=K1+1}^{s} \gamma_{T+k}(f_{T+k|T}(x) - f_x(x))| \\
&\leq |\sum_{k=1}^{K1} 1/s(f_{T+k|T}(x) - f_x(x))| + \sum_{k=K1+1}^{s} \gamma_{T+k}|(f_{T+k|T}(x) - f_x(x))| \\
&\leq 1/s\frac{K1 + K2}{2}\delta + \sum_{k=K1+1}^{s} \gamma_{T+k}\delta/2 \\
&\leq \frac{K1 + K2}{2s}\delta + \frac{1}{2}\delta \\
&< \delta. \qquad (126)
\end{aligned}$$

Thus, $\lim_{s\to\infty} f_I = f_x$. The condition placed upon the sequence $\{\gamma_{T+k}\}$ is a sufficient condition for the proof of the convergence of f_I to f_x and explicitly provides for the two intuitive weighting schemes: equal weighting and exponential weighting.

The density function f_x corresponds to the long-term capability. When the period of interest is long enough, the capability of the process is best described as a function of the limiting or unconditional distribution of the observations. *Long-term process capability*, denoted $\mathcal{C}$, is defined as the ability of the process to produce items within its specification limits at some arbitrarily distant point in time. Using the previously developed notation, the long-term capability can be expressed as

$$\mathcal{C} \triangleq P[LSL \leq X \leq USL]. \qquad (127)$$

Similarly, the *long-term expected value* of the process can be expressed as

$$\mu_x \triangleq E[X] \tag{128}$$

and the *long-term variance* of the process can be expressed via

$$\sigma_x^2 \triangleq Var[X]. \tag{129}$$

The notation introduced so far in this chapter recognizes the possibility that, given a set of known observations, the conditional expected process location at some future time may differ from the time-average expected process location during the period of interest or the unconditional process location. Similarly, process spread, as measured by the conditional variance of the observations, may also change with time. A simple case showing these differences is presented in section 4.4. These differences are important because they directly lead to the ambiguity that is present in the current, static measures of capability.

4.2.3 Redefining Some Common Capability Indices. Capability indices are a preferred method for reporting the capability of a process. The two most generally accepted capability indices are C_{pk} and C_{pm}. These two key process capability indices can be expressed as the long-term capability indices

$$\begin{aligned} \mathcal{C}_{pk} &= min\left[\frac{USL-\mu_x}{3\sigma_x}, \frac{\mu_x - LSL}{3\sigma_x}\right] \\ &\text{and} \\ \mathcal{C}_{pm} &= \frac{USL-LSL}{6\sqrt{(\mu_x-\tau)^2+\sigma_x^2}}. \end{aligned} \tag{130}$$

At some future time, $T+k$, the corresponding time-specific capability indices given conditional knowledge about the observations up to time T are

$$\begin{aligned} \mathcal{C}_{pk,T+k|T} &= min\left[\frac{USL-\mu_{T+k|T}}{3\sigma_{T+k|T}}, \frac{\mu_{T+k|T}-LSL}{3\sigma_{T+k|T}}\right] \\ &\text{and} \\ \mathcal{C}_{pm,T+k|T} &= \frac{USL-LSL}{6\sqrt{(\mu_{T+k|T}-\tau)^2+\sigma^2_{T+k|T}}}. \end{aligned} \tag{131}$$

Finally, when a period of interest for the process is known, the time-average capability indices for the period given the observations up to time T are

$$\begin{aligned} \mathcal{C}_{pk,I} &= min\left[\frac{USL-\mu_I}{3\sigma_I}, \frac{\mu_I-LSL}{3\sigma_I}\right] \\ &\text{and} \\ \mathcal{C}_{pm,I} &= \frac{USL-LSL}{6\sqrt{(\mu_I-\tau)^2+\sigma^2_I}}. \end{aligned} \tag{132}$$

It is important to note that the time-average capability indices are not necessarily equal to the average of the time-specific capability indices over the period of interest. However, time-average capability indices do provide a measure of the total variation exhibited by the process throughout the period of interest that is not unduly influenced by the variation at isolated times in the period of the interest.

The multifaceted nature of capability is recognized and accounted for by this notation. Time-average capability is a measure of the ability of a process to produce within specification limits throughout the entire period of interest. On the other hand, time-specific capability is a measure of the ability of the process to produce within the specification limits *at a given future point in time.* Finally, long-term capability is an unconditional measure of the ability of the process to produce within the specification limits.

4.2.4 Time-Specific, Time-average and Long-Term Expected Loss Defined. The major alternative to considering capability in terms of the probability of observations being within the specification limits is to consider capability in terms of the expected loss per observation. The measure of expected loss per observation possesses the same three aspects associated with the probability based measure of capability. Expected loss is frequently approximated by the Taguchi Loss Function, $L(x) = K(x - \tau)^2$ where K is a constant. Similar to long-term capability, the long-term expected loss per observation, denoted $\mathcal{L}$, given a specified Taguchi loss function can be written as:

$$\begin{aligned}
\mathcal{L} &= E[L(X)] \\
&= E[K(X-\tau)^2] \\
&= K(E[X^2] - E[2\tau X] + E[\tau^2]) \\
&= K(Var[X] + E^2[X] - 2\tau\ E[X] + \tau^2) \\
&= K(Var[X] + (E[X] - \tau)^2) \\
&= K(\sigma_x^2 + (\mu_x - \tau)^2). \qquad (133)
\end{aligned}$$

The time-specific expected loss per observation at time $T+k$ given observations up to time T follows, via

$$\begin{aligned}
\mathcal{L}_{T+k|T} &= E[L(X_{T+k})|x_1, \ldots, x_T] \\
&= K(\sigma_{T+k|T}^2 + (\mu_{T+k|T} - \tau)^2), \qquad (134)
\end{aligned}$$

and the time-average expected loss per observation during the period of interest via

$$\begin{aligned}
\mathcal{L}_I &= E[L(X_I)|x_1, \ldots, x_T] \\
&= K(\sigma_I^2 + (\mu_I - \tau)^2). \qquad (135)
\end{aligned}$$

The time-average expected loss per observation during the period of interest is equal to the weighted average of the time-specific expected loss per observation.

Boyles (1991) describes an inverse square relationship between the (long-term) C_{pm} and the Taguchi loss function. When

$$K = 36/(USL - LSL)^2 \tag{136}$$

then

$$\mathcal{L} = \frac{1}{\mathcal{C}_{pm}^2}. \tag{137}$$

That relationship is preserved under our definitions of time-average and time-specific capability and loss, via

$$\begin{aligned} \mathcal{L}_I &= \frac{1}{\mathcal{C}_{pm,I}^2} \\ \mathcal{L}_{T+k|T} &= \frac{1}{\mathcal{C}_{pm,T+k|T}^2}. \end{aligned} \tag{138}$$

4.3 Independent and Identically Distributed Case.

The expanded definitions of capability and expected loss per unit address issues raised by applying the existing definitions to modern processes which exhibit autocorrelation. The existing definitions adequately handle independent and identically distributed processes. In this section, we verify that the long-term capability, time-average capability and time-specific capability are all equal when process observations are independent and identically distributed.

Suppose we have independent and identically distributed process observations, denoted by the random variable X. The presumed truth model for this case is

$$X_t = \mu_x + \epsilon_t \tag{139}$$

where μ_x is a constant and ϵ_t is a random noise component such that

$$\begin{aligned} E(\epsilon_t) &= 0 \quad \forall\, t \\ Var(\epsilon_t) &= \sigma_\epsilon^2 \quad \forall\, t \\ Cov(\epsilon_t, \epsilon_{t+k}) &= 0 \quad \forall\, t,\ \forall\, k > 0. \end{aligned} \tag{140}$$

The noise component is frequently assumed to follow a normal distribution, although that assumption is not relied upon in this chapter. It is easily seen that the following relations are true for this case:

$$\begin{aligned} E(X_t) &= \mu_x \quad \forall\, t \\ Var(X_t) &= \sigma_\epsilon^2 \quad \forall\, t \\ Cov(X_t, X_{t+k}) &= 0 \quad \forall\, t,\ \forall\, k > 0 \\ \sigma_x^2 &= \sigma_\epsilon^2. \end{aligned} \tag{141}$$

It is also straightforward from the independence of the process that future observations do not rely upon previous observations. For instance, suppose observations up to time T are known. The relations from above still hold given the conditional information, as

$$\begin{aligned} E(X_{T+k}|x_1, \ldots, x_N) &= \mu_x \quad \forall\, t \geq T+1 \\ Var(X_{T+k}|x_1, \ldots, x_N) &= \sigma_\epsilon^2 \quad \forall\, t \geq T+1 \\ Cov(X_{T+l}, X_{T+k}|x_1, \ldots, x_N) &= 0 \quad \forall\, l > 0,\ \forall\, k > 0. \end{aligned} \tag{142}$$

The time invariance of the independent and identically distributed case extends to the capability of the process. Knowledge about past observations does not affect expectations about future events, so

$$P[LSL \leq X_{T+k} \leq USL | x_1, \ldots, x_N] = P[LSL \leq X \leq USL] \quad \forall\, k \geq 0, \tag{143}$$

and in general, given the observations up to time T, for any $k \geq 1$, the long-term, time-average and time-specific capability indices are equal, via

$$\begin{aligned} \mathcal{C}_{T+k|T} &= \mathcal{C} = & \mathcal{C}_I \\ \mathcal{C}_{pk,T+k|T} &= \mathcal{C}_{pk} = & \mathcal{C}_{pk,I} \\ \mathcal{C}_{pm,T+k|T} &= \mathcal{C}_{pm} = & \mathcal{C}_{pm,I}. \end{aligned} \tag{144}$$

In other words, when the observations are independent and identically distributed, there is no ambiguity when using static measures of capability. In the following sections, we show how ambiguity arises when the observations are not independent and identically distributed.

4.4 Deterministic Case

Another simple case can be used to show that the long-term and time-specific capabilities are not necessarily equal. A deterministic process shows this quite easily. For instance, consider a sinusoidal process that generates the repeating sequence of observations {0, .7, 1, .7, 0, -.7, -1, -.7, 0, ...}. Further suppose that the upper specification limit for the process is .8 and the lower specification limit is -.8. This process is depicted in Figure 11. It is easy to see that two out of every eight observations will be outside of the specification limits. Thus, the long-term capability as measured by $\mathcal{C}$ equals .75. However, at any given time the process is either inside or outside of the specification limits. Therefore, $\mathcal{C}_{T+k|T}$ will equal either 0 or 1 and $\mathcal{C}_{T+k|T} \neq \mathcal{C}$. Note that for a deterministic seasonal process like this one, it is not correct to say that $f_{T+k|T}$ approaches a limiting distribution. However, we can say that the linear combination of probability distribution

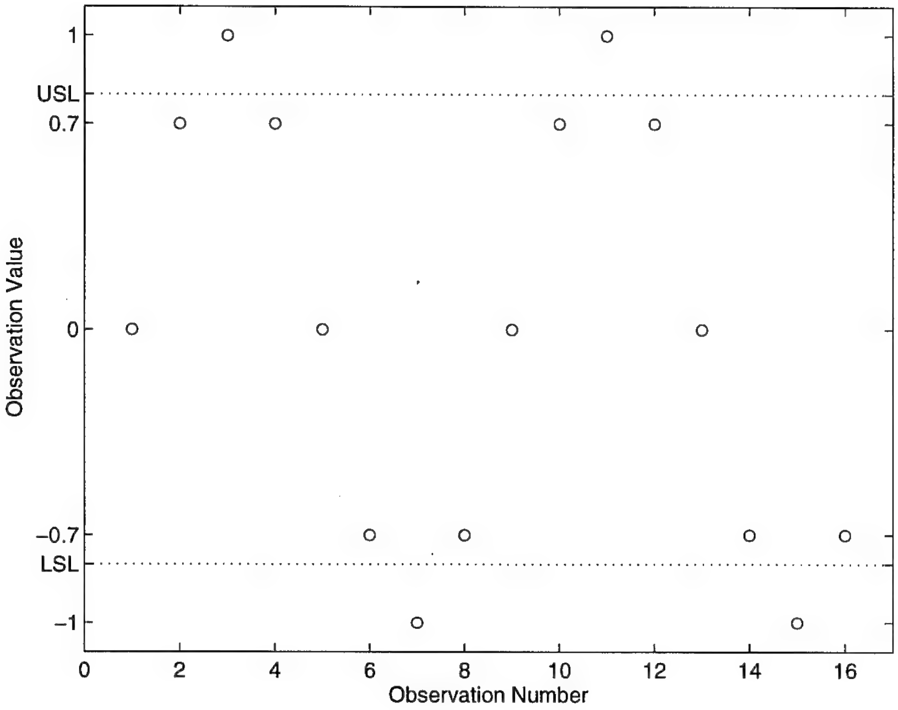


Figure 11. Example of a sinusoidal process.

functions over the season approaches a limiting function, and so, f_x still does approach a limiting function.

4.5 General Case

As discussed in Chapter II, the independence assumption does not hold for a large variety of real world processes. The more general case allows for future process observations to depend upon past process observations as well as past values of the error component. In this case, the truth model can be expressed as

$$X_t = f(x_1, \ldots, x_{t-1}, \epsilon_1, \ldots, \epsilon_{t-1}) + \epsilon_t \tag{145}$$

where f is some function of past process observations and errors. For example, an ARMA(1,1) model can be defined by

$$f(x_1, \ldots, x_{t-1}, \epsilon_1, \ldots, \epsilon_{t-1}) = \begin{cases} \xi + \phi x_0 - \theta \epsilon_0 & \text{for } t = 1 \\ \xi + \phi x_{t-1} - \theta \epsilon_{t-1} & \text{for } t \geq 2 \end{cases} \tag{146}$$

where ξ, ϕ and θ are the ARMA model parameters; and x_0 and ϵ_0 account for the model's initial conditions.. A slightly more general formulation of the truth model would include ϵ_t as a variable in the function, but it will be convenient later to assume the error term is added at each time step.

Suppose the N most recent observations at time T from a general process that can be described by equation 145 are known. We already know that the expected value of X at time $T+k$ given the observations up to time T, denoted $\mu_{T+k|T}$, is not necessarily equal to the long-term expected value of the process observations,

$$\begin{aligned} \mu_{T+k|T} &\triangleq E(X_{T+k}|x_1, \ldots, x_N) \\ \mu_{T+k|T} &\neq \mu_x. \end{aligned} \tag{147}$$

Nor will the conditional process variance, denoted $\sigma^2_{T+k|T}$, necessarily equal σ^2_x,

$$\begin{aligned} \sigma^2_{T+k|T} &\triangleq Var(X_{T+k}|x_1, \ldots, x_N) \\ \sigma^2_{T+k|T} &\neq \sigma^2_x. \end{aligned} \tag{148}$$

The long-term capability of the general process can be assessed by determining values for $\mathcal{C}$, $\mathcal{L}$, $\mathcal{C}_{pk}$ or $\mathcal{C}_{pm}$, just as the long-term capability can be assessed for the independent and identically distributed case. However, the time-specific capability of the process in the more general case does not necessarily equal the long term capability of the process. That is,

$$P[LSL \leq X_{T+k} \leq USL | x_1, \ldots, x_T] \neq P[LSL \leq X \leq USL]. \tag{149}$$

4.5.1 An Example. Suppose we have an AR(1) process with autoregressive parameter $\phi = .9$, standard normal errors, and a mean of zero. The process is characterized by

$$X_{t+1} = \phi X_t + \epsilon_t. \tag{150}$$

The long-term standard deviation of the process is given by

$$\begin{aligned} \sigma_x &= 1/(1-\phi^2) \\ &= 5.26. \end{aligned} \tag{151}$$

Further suppose that the specification limits are set at $\pm 3\sigma_x = \pm 15.78$. Then, the long-term capability of the process can be expressed by

$$\begin{aligned} \mathcal{C}_{pk} &= \min\left[\frac{USL-0}{3\sigma_x}, \frac{0-LSL}{3\sigma_x}\right] \\ &= 1. \end{aligned} \tag{152}$$

In addition, suppose that the process observation measured at time $t = 1$ was 15.8. Then, the conditional expected value of the second observation is

$$\begin{aligned} \mu_{2|1} &= E[X_2 | x_1 = 15.8] \\ &= E[.9x_1 + \epsilon_2 | x_1 = 15.8] \\ &= .9(15.8) + 0 \\ &= 14.2 \end{aligned} \tag{153}$$

and the conditional variance is

$$\begin{aligned} \sigma^2_{2|1} &= Var[X_2|x_1 = 15.8] \\ &= Var[.9x_1 + \epsilon_2|x_1 = 15.8] \\ &= Var[\epsilon_2] \\ &= 1. \end{aligned} \tag{154}$$

Therefore, the conditional time-specific capability of the process at time $t = 2$ is given by

$$\begin{aligned} \mathcal{C}_{pk,2|1} &= \min\left[\frac{USL - \mu_{2|1}}{3\sigma_{2|1}}, \frac{\mu_{x,2|1} - LSL}{3\sigma_{x,2|1}}\right] \\ &= \min[.531, 10.0] \\ &= .531. \end{aligned} \tag{155}$$

This example shows that there can be quite a large difference between the long-term capability, $\mathcal{C}_{pk} = 1$, and the time-specific capability, $\mathcal{C}_{pk,2|1} = .531$. If our measurement of the observation at time $t = 2$ is 0 (a highly unlikely event given the truth model), then $\mathcal{C}_{pk,3|2} = 5.26$; the time-specific process capability can be much larger than the long-term process capability.

4.6 One-Step Ahead Process Capability.

In the preceding sections, we demonstrated the need for considering time varying capability. In the following sections, we will develop techniques for examining time varying capability. Consider the natural question, what is the time-specific capability of the process for the next observation? That is, when the N most recent observations at time T are known, what is the probability that the observation at time $T + 1$ will be within the specification limits? The one-step ahead capability at time T is given by

$$\mathcal{C}_{T+1|T} \triangleq P(LSL \leq X_{T+1} \leq USL|x_1, \ldots, x_T). \tag{156}$$

The one step ahead capability can be determined directly if the conditional distribution of X_{T+1} is known. More generally, X_{T+1} can be assumed to arise from a known family of models. By fitting a general model to the N known observations, we are able to estimate $\mathcal{C}_{T+1|T}$ via

$$\hat{\mathcal{C}}_{T+1|T} \triangleq P(LSL \leq \hat{X}_{T+1} \leq USL|x_1, \ldots, x_T). \tag{157}$$

4.6.1 General Case. For the general case, X_{T+1} can be expressed using equation 145 via

$$X_{T+1} = f(x_1, \ldots, x_N, \epsilon_1, \ldots, \epsilon_T) + \epsilon_{T+1}. \tag{158}$$

Unfortunately, neither the parameters of the function f nor the distribution of the error component are generally known. However, an estimator for the one step ahead capability can be constructed using estimates of the parameters of the truth model. Suppose a candidate function, g, is being considered as an approximation for the truth model described by the function f. Since an observation is known for each of the first T time periods, the relationship

$$x_i = g(x_1, \ldots, x_{i-1}, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_{i-1}) + \hat{\epsilon}_i, \quad (1 \leq i \leq T) \tag{159}$$

can be established with some estimates, $\{\hat{\epsilon}\}$, replacing the true error components, $\{\epsilon\}$. (In the forecasting literature, e is frequently used in place of $\hat{\epsilon}$.) A commonly used technique is to choose parameters for g which minimize the sum of the square of the estimated errors. That is, the parameters of function g can be fitted by solving the minimization problem:

$$\begin{aligned} \text{minimize}: \quad & \sum_{i=1}^{T} \hat{\epsilon}_i^2 \\ \text{subject to}: \quad & x_i = g(x_1, \ldots, x_{i-1}, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_{i-1}) + \hat{\epsilon}_i \ \ \forall\, 1 \leq i \leq T \end{aligned} \tag{160}$$

The solution of the minimization problem yields the parameters for g and a set of estimated error components, $\{\hat{\epsilon}\}$. Then, $g(x_1, \ldots, x_T, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_T)$ can be used as an estimate of $f(x_1, \ldots, x_T, \epsilon_1, \ldots, \epsilon_T)$. The expected value of the next process observation is given by

$$\mu_{T+1|T} = E[f(x_1, \ldots, x_T, \epsilon_1, \ldots, \epsilon_T) + \epsilon_{T+1} | x_1, \ldots, x_N] \tag{161}$$

and can be estimated via

$$\begin{aligned}\hat{\mu}_{T+1|T} &= E[g(x_1, \ldots, x_T, \epsilon_1, \ldots, \epsilon_T) + \epsilon_{T+1} | x_1, \ldots, x_N] \\ &= E[g(x_1, \ldots, x_T, \epsilon_1, \ldots, \epsilon_T) | x_1, \ldots, x_N] + E[\epsilon_{T+1} | x_1, \ldots, x_N] \\ &= g(x_1, \ldots, x_T, \hat{\epsilon}_1, \ldots, \hat{\epsilon}_T). \end{aligned} \tag{162}$$

Using this new information, $\hat{\mathcal{C}}_{T+1|T}$ can be expressed via

$$\begin{aligned}\hat{\mathcal{C}}_{T+1|T} &= P(LSL \leq \hat{X}_{T+1} \leq USL) \\ &= P(LSL - \hat{\mu}_{T+1|T} \leq \epsilon_{T+1} \leq USL - \hat{\mu}_{T+1|T}). \end{aligned} \tag{163}$$

As would be expected, the one step ahead process capability can be estimated using estimates of the one step ahead process location and the variance of the errors. An estimate of the one step ahead process location can be directly obtained from the fitted model and errors. However, in order to derive a value for the estimated capability using equation 163, some assumptions must be made about the distribution of ϵ_{T+1}. In the next chapter, we propose a method for doing so.

4.6.2 Independent and Identically Distributed Case. Suppose that the candidate model for the process is chosen from the family of independent and identically distributed processes with an unknown mean. That is, choose $g(\cdot) = K$. Then the minimization problem

$$\begin{aligned} \text{minimize:} \quad & \sum_{i=1}^{T} \hat{\epsilon}_i^2 \\ \text{subject to:} \quad & x_i = K + \hat{\epsilon}_i \ \ \forall\, 1 \leq i \leq T \end{aligned} \tag{164}$$

simplifies to

$$\text{minimize:} \sum_{i=1}^{T} (x_i - K)^2 \tag{165}$$

with the solution

$$K = \sum_{i=1}^{T} x_i / T \tag{166}$$

and yields the set of estimated errors $\{x_i - K\}$. For this process, the one-step ahead capability is given by

$$\mathcal{C}_{T+1|T} = P(LSL - \mu_x \leq \epsilon_{T+1} \leq USL - \mu_x). \tag{167}$$

The estimated one-step ahead capability for this case is given by

$$\begin{aligned} \hat{\mathcal{C}}_{T+1|T} &= P(LSL - \hat{\mu}_{T+1|T} \leq \epsilon_{T+1} \leq USL - \hat{\mu}_{T+1|T}) \\ &= P(LSL - \sum_{i=1}^{T} x_i/T \leq \hat{\epsilon}_{T+1} \leq USL - \sum_{i=1}^{T} x_i/T). \end{aligned} \tag{168}$$

It is clear that as the number of observations increase, the estimate of K will approach μ_x, via

$$\begin{aligned} \lim_{T\to\infty} K &= \lim_{T\to\infty} \sum_{i=1}^{T} x_i/T \\ &= \mu_x. \end{aligned} \tag{169}$$

Further, the set of estimated errors approach the true errors, since,

$$\begin{aligned} \lim_{T\to\infty} \hat{\epsilon}_i &= \lim_{T\to\infty} (x_i - K) \\ &= x_i - \mu_x \\ &= \epsilon_i. \end{aligned} \tag{170}$$

Finally, the estimated one step ahead capability estimate approaches the true long-term capability, or,

$$\lim_{T\to\infty} \hat{\mathcal{C}}_{T+1|T} = \mathcal{C}. \tag{171}$$

This is an intuitively appealing result. Given a large enough sample of independent and identically distributed observations, the probability density function of the observations and, hence, of the error terms, can be approximated to an arbitrary degree of accuracy. Process capability can, in turn, be approximated from that unconditional distribution.

4.6.3 Deterministic Case. The deterministic case retains its intuitive simplicity when estimating the one step ahead capability. When the deterministic pattern is known, the one step ahead forecast errors are uniformly equal to zero. In addition, the location of the process at the next time step is known. Therefore, the one step ahead capability is given by

$$\begin{aligned} \mathcal{C}_{T+1|T} &= P(LSL - \mu_{T+1|T} \leq \epsilon_{T+1} \leq USL - \mu_{T+1|T}) \\ &= P(LSL - \mu_{T+1|T} \leq 0 \leq USL - \mu_{T+1|T}) \\ &= P(LSL \leq \mu_{T+1|T} \leq USL). \end{aligned} \tag{172}$$

Since the location of the process at the next time step is known, the one step ahead capability will equal either 0 (if the location is outside of the specification limits) or 1 (if the location is inside of the specification limits).

4.7 k-Step Ahead Process Capability.

In this section, we generalize the results from the one-step ahead capability section. Specifically, we modify the equations for assessing one-step ahead process capability to assess the k-step ahead process capability.

4.7.1 General Case. Following the same development used in the one-step ahead case,

$$\begin{aligned}\hat{\mu}_{T+k|T} &= E[\hat{X}_{T+k}|x_1,\ldots,x_T] \\ &= E[g(x_1,\ldots,x_{T+k-1},\hat{\epsilon}_1,\ldots,\hat{\epsilon}_{T+k-1})+\epsilon_{T+k}|x_1,\ldots,x_T] \\ &= E[g(x_1,\ldots,x_{T+k-1},\hat{\epsilon}_1,\ldots,\hat{\epsilon}_{T+k-1})|x_1,\ldots,x_T]+E[\epsilon_{T+k}|x_1,\ldots,x_T] \\ &= g^k(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T), \qquad (173)\end{aligned}$$

where $g^k(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T)$ is recursively defined by

$$\begin{aligned}g^k(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T) &= g(x_1,\ldots,x_T,g^1(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T),\ldots, \\ &\qquad g^{k-1}(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T),\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T,0,\ldots,0) \\ &\ \vdots \\ g^3(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T) &= g(x_1,\ldots,x_T,g^1(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T), \\ &\qquad g^2(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T),\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T,0,0) \\ g^2(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T) &= g(x_1,\ldots,x_T,g^1(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T),\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T,0) \\ g^1(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T) &= g(x_1,\ldots,x_T,\hat{\epsilon}_1,\ldots,\hat{\epsilon}_T). \qquad (174)\end{aligned}$$

That is, the forecasting equation is successively evaluated for each time step forward using previous forecasts along with an expected error of zero. Then, the k-step ahead process capability becomes

$$\begin{aligned}\mathcal{C}_{T+k|T} &= P(LSL \leq X_{T+k} \leq USL|x_1, \ldots, x_T) \\ &= P(LSL - \mu_{T+k|T} \leq \epsilon_{T+k} \leq USL - \mu_{T+k|T}) \qquad (175)\end{aligned}$$

and is estimated via

$$\hat{\mathcal{C}}_{T+k|T} = P(LSL - \hat{\mu}_{T+k|T} \leq \epsilon_{T+k} \leq USL - \hat{\mu}_{T+k|T}). \qquad (176)$$

4.7.2 Independent and Identically Distributed Case. For the independent and identically distributed case, the k-step ahead capability estimate is equal to the one-step ahead capability estimate. This is an intuitive result since the process location and process spread do not change. Indeed, since $g(\cdot)$ is defined as a constant K, it is clear that $g^k(\cdot)$ is also equal to the same constant. Furthermore, ϵ_{T+k} will have the same distribution as ϵ_{T+1}. Therefore

$$\begin{aligned}\hat{\mathcal{C}}_{T+k|T} &= P(LSL - \hat{\mu}_{x,T+k|T} \leq \epsilon_{T+k} \leq USL - \hat{\mu}_{x,T+k|T}) \\ &= P(LSL - \hat{\mu}_{x,T+1|T} \leq \epsilon_{T+1} \leq USL - \hat{\mu}_{x,T+1|T}) \\ &= \hat{\mathcal{C}}_{T+1|T}. \qquad (177)\end{aligned}$$

4.7.3 Deterministic Case. The deterministic case is also straightforward. Successive application of the fitted model yields an estimate of the process location k time-steps forward. Since the sequential pattern of observations is known, there is no variance about the process location. Therefore, as with the one-step ahead estimate, the k-step ahead estimate of capability will equal either 0 or 1.

4.8 Other Measures of Capability.

The techniques and notation used to describe the k-step ahead process capability can be directly applied to other measures of process capability. The $\mathcal{C}_{pk}$ process capability index statically describes the overall process capability. Using the time-varying techniques proposed in this paper, the k-step ahead $\mathcal{C}_{pk}$ index can be defined

$$\mathcal{C}_{pk,T+k|T} = min\left[\frac{USL - \mu_{T+k|T}}{3\sigma_{T+k|T}}, \frac{\mu_{T+k|T} - LSL}{3\sigma_{T+k|T}}\right] \tag{178}$$

and can then be estimated via

$$\hat{\mathcal{C}}_{pk,T+k|T} = min\left[\frac{USL - \hat{\mu}_{T+k|T}}{3\hat{\sigma}_{T+k|T}}, \frac{\hat{\mu}_{T+k|T} - LSL}{3\hat{\sigma}_{T+k|T}}\right]. \tag{179}$$

Similarly, the k-step ahead $\mathcal{C}_{pm}$ index can be defined via

$$\mathcal{C}_{pm,T+k|T} = \frac{USL - LSL}{6\sqrt{(\mu_{T+k|T} - \tau)^2 + \sigma^2_{T+k|T}}} \tag{180}$$

and estimated via

$$\hat{\mathcal{C}}_{pm,T+k|T} = \frac{USL - LSL}{6\sqrt{(\hat{\mu}_{T+k|T} - \tau)^2 + \hat{\sigma}^2_{T+k|T}}}. \tag{181}$$

Finally, the k-step ahead expected loss per observation is given by

$$\mathcal{L}_{T+k|T} = K[\sigma^2_{T+k|T} + (\mu_{T+k|T} - \tau)^2] \tag{182}$$

and is estimated by

$$\hat{\mathcal{L}}_{T+k|T} = K[\hat{\sigma}^2_{T+k|T} + (\hat{\mu}_{T+k|T} - \tau)^2]. \tag{183}$$

4.9 Chapter Summary.

The concept of capability is generally understood by quality practitioners. However, the static measures currently used to measure capability do not adequately address the

Table 12. Summary of measures of capability for different time frames.

Basis for Measuring Capability	Time-Frame					
	Long-term		Average		Time-specific	
	Measure	Estimate	Measure	Estimate	Measure	Estimate
Probability Based	$\mathcal{C}$	$\hat{\mathcal{C}}$	$\mathcal{C}_I$	$\hat{\mathcal{C}}_I$	$\mathcal{C}_{T+k\vert T}$	$\hat{\mathcal{C}}_{T+k\vert T}$
Loss Based	$\mathcal{L}$	$\hat{\mathcal{L}}$	$\mathcal{L}_I$	$\hat{\mathcal{L}}_I$	$\mathcal{L}_{T+k\vert T}$	$\hat{\mathcal{L}}_{T+k\vert T}$
C_{pk} Based	$\mathcal{C}_{pk}$	$\hat{\mathcal{C}}_{pk}$	$\mathcal{C}_{pk,I}$	$\hat{\mathcal{C}}_{pk,I}$	$\mathcal{C}_{pk,T+k\vert T}$	$\hat{\mathcal{C}}_{pk,T+k\vert T}$
C_{pm} Based	$\mathcal{C}_{pm}$	$\hat{\mathcal{C}}_{pm}$	$\mathcal{C}_{pm,I}$	$\hat{\mathcal{C}}_{pm,I}$	$\mathcal{C}_{pm,T+k\vert T}$	$\hat{\mathcal{C}}_{pm,T+k\vert T}$

time varying aspects of capability required by modern processes which exhibit autocorrelation. The explicit definitions for time-specific, time-average and long-term capability we presented in this chapter fill a void in the current literature by addressing the issue of how to define capability for autocorrelated processes. We presented numerous examples in this chapter to demonstrate the need for time varying measures of capability. The time varying definitions we presented in this chapter lead to a diverse set of measures of capability. Table 12 summarizes these measures. We also showed how time-specific, time-average and long-term capability can be estimated by fitting a candidate model to known process observations. The concepts we developed in this chapter provide the foundation for the development of a capability monitoring system in the next chapter.

V. A System for Monitoring Process Capability.

Taguchi (1985) considers the two key problems in quality improvement to be how to measure quality and how to improve quality. One way to measure process quality is to quantify process capability. Traditionally, capability indices provide static estimates of process capability. In this research, we propose considering capability as a time-varying aspect of the process. Looked at in this way, process capability can be monitored in the same manner as statistical control and has the added benefit of directly tracking a measure of quality. In the previous chapter, we presented the mathematical foundation for predicting process capability. In this chapter, we use that foundation to develop a practical method for monitoring process capability.

The objective of this chapter is to demonstrate the potential value of a capability based monitoring system. In the first section of this chapter, we discuss some general goals for a capability monitoring system. That discussion is followed by our description of a general capability monitoring system. The basis for the capability monitoring system is a statistical test to determine whether a process lacks capability at a given time. We discuss the theory and definitions underlying that statistical test. Then, we present a proposal for a specific capability monitoring system. We test the proposed system by Monte Carlo simulation for a variety of ARMA(1,1) models. Finally, we explore the added value gained by knowing the exact parameters for the truth model.

5.1 Goals for a Capability Monitoring Method

For much of the past century, the Shewhart control chart has proven to be a practical method for monitoring the state of statistical control. It was successfully applied to industrial processes which often lacked even the most basic control over the quality of output. That control, provided by the Shewhart control chart, is required as one of the first steps toward quality improvement. To that end, the Shewhart control chart has met the needs of the people who have used it. The Shewhart control chart worked best when the observations it was applied to could be reasonably modeled as independent and identically distributed. For such processes, the Shewhart control charts were able to detect the kinds of assignable causes of variation that were occurring in the processes and, thus, assisted in the identification and removal of those causes. Finally, the Shewhart control chart was appropriate to the level of computational power available in the field. It only required simple mathematics that could be done by hand.

The objective of the capability monitoring system proposed in this chapter is to satisfy the quality improvement requirements for a broader class of processes, including processes that exhibit autocorrelation. While keeping a process in a state of statistical control remains an important consideration, an increased emphasis is being placed upon process capability. We propose direct monitoring of process capability. A limitation of the Shewhart control charts is its unpredictable performance when process observations are autocorrelated. A practical monitoring system should be applicable to a wide variety of processes, including processes that generate autocorrelated observations. When applied to a process that has both chance and structural cause variation, the system should have a predictable average run length in the absence of assignable causes of variation, yet should be able to detect to the addition of an assignable cause of variation. The computers found in today's workplace are an increasingly powerful tool for quality improvement. A capability monitoring system should be allowed to take advantage of that power, and should provide for frequent on-line evaluation of observations. The goals for the capability monitoring system proposed in this chapter can be summarized as:

- Monitor the capability of the process.
- Respond to a variety of assignable causes, including a shift in the process mean.
- Apply to any process from the family of stationary ARMA(1,1) models.

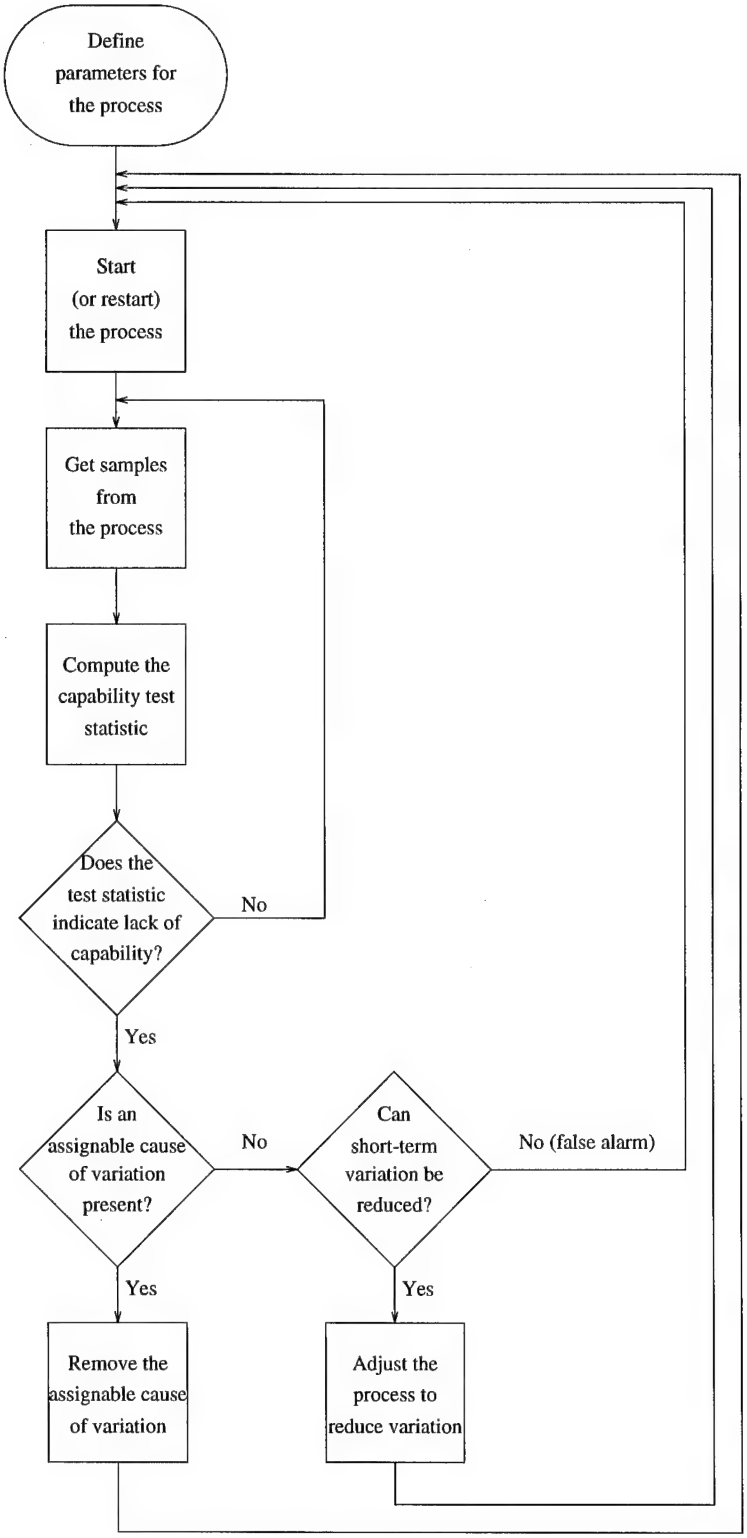
- Be computationally efficient; able to be implemented on-line using readily available hardware platforms (e.g., within times that are small relative to the sampling interval).

5.2 A General Method for Monitoring Process Capability

Monitoring process capability can be considered an application of statistical thinking. The primary benefit of monitoring process capability is the detection and identification of assignable causes of variation, leading to the elimination of those assignable causes. Statistical information about the process gained coincidentally by a capability monitoring system can also aide in identifying potential system changes that could reduce or eliminate chance or structural causes of variation. The relationship between statistical thinking and capability monitoring is analogous to the relationship between statistical thinking and monitoring the state of statistical control.

A flowchart for a general capability monitoring system is depicted in Figure 12. Each block in the figure represents a logical step in the system and the connecting arrows identify the flow of steps taken. Each step in the general capability monitoring system depicted in Figure 12 is described in this section. This general system described in this section will be expanded in following sections into our proposed system for monitoring process capability.

5.2.1 Parameter Definition. A number of parameters must be specified for any process monitoring system. The topmost block in the flowchart represents the definition step of the process. At the most basic level, we must know which quality characteristic of the process is to be measured and the range of allowable values for the measurements. For the system proposed in this chapter, and for most others, the process measurements are assumed to be real-valued observations taken at fixed intervals. In addition to defining the process, some quality related parameters must be specified in order to measure the



Define parameters for the process
Start (or restart) the process
Get samples from the process
Compute the capability test statistic
Does the test statistic indicate lack of capability?
No
Yes
Is an assignable cause of variation present?
No
Can short-term variation be reduced?
No (false alarm)
Yes
Yes
Remove the assignable cause of variation
Adjust the process to reduce variation

Figure 12. Flowchart of a general method for monitoring process capability.

capability of the process. For instance, upper and lower specification limits and a target value for the process measurements allow an estimate of process capability to be made.

5.2.2 Process Start-up. After any necessary parameters are defined, the process can be started. The second block in the flowchart represents this step. At this step, the process may be physically started or 'reset' to its nominal configuration. In addition, depending on the implementation of the system, some variables may need to be reset to their starting values.

Three additional arrows point into this block. The first alternative path invokes the initialization step after the removal of an assignable cause of variation. In this case, the process may or may not have been temporarily stopped during the search for and removal of the assignable cause. When an assignable cause is identified and removed, it is likely that some re-initialization of variables will be required, although the parameters defined in the top block should remain valid. The second alternative path to enter this block follows a process adjustment to temporarily reduce the effects of structural cause variation. In this case, the process can be considered to be re-started. Finally, this block may be entered after the indicated lack of capability is deemed to be a false alarm. The process may need to be re-started due to time spent in evaluating the signalled lack of capability.

5.2.3 Sampling from the Process. The sampling strategy plays an important role in defining the system. The sampling strategy may include determining sample sizes, the frequency with which observations are to be taken, and any transformations to be applied to the observed quality measurements. The samples are used to compute an appropriate test statistic.

5.2.4 Compute the Capability Test Statistic. The test statistic is the major difference between a capability monitoring system and a control monitoring system. In a capability monitoring system, the test statistic is assumed to provide some measure of the true capability of the process. In a control monitoring system, the test statistic provides a measure of state of statistical control indirectly by measuring some aspect of the distribution of the observations (e.g. the mean or variance of the observations). In either case, the distribution of the test statistic should change following the introduction of an assignable cause of variation.

5.2.5 Perform the Capability Test. A test is performed using the test statistic which leads to a decision to either accept or reject the hypothesis that the process is capable. The subset of the range of the test statistic which, in accordance with the capability test, leads to the rejection of the capability hypothesis is called the critical region of the test. The capability hypothesis will be discussed in detail in Section 5.3. When the test statistic falls within the critical region, the hypothesis of capability is rejected. In this case, an assignable cause of variation may have arisen, so the flow through the system passes down to search for an assignable cause. If the test statistic does not fall in the critical region, then insufficient evidence for the lack of capability exists to reject the hypothesis of capability. In this case, no remedial action is called for and the flow through the system loops back up to continue sampling.

5.2.6 Search for an Assignable Cause. Although this step is only invoked when statistical evidence of the lack of capability exists, that evidence does not necessarily mean an assignable cause of variation is present. Since the test statistic is a random variable, some false alarms are expected. The search for an assignable cause of variation generally requires a human interpretation of conditions which may have affected the process. The search may be facilitated by examining a time-history of the test statistic and other process statistics, such as local mean and variance estimates. If an assignable cause of variation is found, the flow through the general system proceeds down to the removal step. In control monitoring systems, when an assignable cause of variation is not found, we assume that the signal from the statistical test is a false alarm. However, since the capability monitoring system explicitly allows for structural cause variation, the capability monitoring system reflects the possibility of proceeding by examining the short-term variation.

5.2.7 Investigate Short-Term Variation. The presence of structural cause variation might be exploited to reduce the short-term variation. An indication of lack of capability might reflect that an autocorrelated process has wandered away from its target value. This step determines whether process changes, such as re-centering the process to the target value are feasible. If changes are feasible, we proceed to adjust the process. If not, we conclude that the signal is truly a 'false alarm.'

5.2.8 Removal of an Assignable Cause. The removal of an assignable cause of variation is a step that is highly specific to both the process and the source of the assignable cause. Like the search for an assignable cause, the removal of an assignable cause generally requires human intervention. After removal, the monitoring system resumes, following any necessary re-initialization.

5.2.9 Adjust the Process to Reduce Short-Term Variation. As with the removal of an assignable cause, this step is highly dependent upon the process being monitored. It is possible that changes to the process to reduce short-term variation simply cannot be made.

5.3 The Capability Hypothesis.

The capability monitoring system we propose in this chapter tests the null hypothesis that the process is capable. We will use the phrase 'capability hypothesis' to refer to this null hypothesis. In this section, we describe the capability hypothesis and the rationale leading to a capability test statistic. The capability hypothesis can be expressed via

$$\begin{gathered} H_0 : \text{The process is capable.} \\ \text{versus} \\ H_A : \text{The process is not capable.} \end{gathered} \tag{184}$$

Our intuitive understanding is that a capable process is a process which will produce almost all of its output within the specification limits. Capability is thus a predictive statement about the process. Recall that Shewhart originally described control in terms of predictability. Later, Wheeler and Chambers concluded that a process must be in control in order for it to be capable. We can logically connect these statements and say that a process must be predictable in order for it to be capable, or, more precisely, a process must be statistically predictable in order for it to be judged capable. Given some knowledge about the chance and structural causes of variation, the location and spread of the process at some time in the future can be estimated. However, the introduction of an assignable cause of variation may result in actual observations quite different from the predicted observations.

According to the Principle of Parsimony (Tukey, 1961): "It may pay not to try to describe in the analysis the complexities that are really present in the situation." On face value, this principle seems to point to accepting the assumption of independence. However, given the difficulties caused by the assumption of independence in the presence of autocorrelation, it appears reasonable to resort to a different base model that can account for the autocorrelation. To rephrase Tukey, it may pay not to try to describe the complexities of the real process with a model beyond a low order ARMA model. In fact, the ARMA(1,1) model appears to be sufficient for the vast majority of processes cited in the literature. The notable exceptions to this claim are those processes with a strong seasonal component (Berthouex et al., 1978). Note that the ARMA(1,1) model encompasses several other common models including the independent normal, AR(1) and MA(1) models. For these reasons, the scope of this research is limited to the ARMA(1,1) family of processes. The ARMA(1,1) process is described in Section 3.2.

One exploitable aspect of the ARMA(1,1) family is the explicit assumption that the underlying error terms, denoted ϵ_t, are independent and identically distributed observations from a normal distribution. When our estimates of future process *properties* are based upon an assumption about the distribution of the error terms, then statistical evidence that the errors do not come from that distribution indicates that the statements about the future *properties* cannot be trusted. When our estimates of future process *capability* are based upon the assumption that the errors driving an ARMA(1,1) process are normally distributed, then statistical evidence that the errors are not normally distributed indicates that the statements about the future *capability* cannot be trusted. As we have already stated, predictability precedes capability. Under the assumption of normally distributed errors, the predictability of an ARMA(1,1) process can be considered to be diminished when the underlying errors are not normally distributed. The capability hypothesis, which will be used for the remainder of this chapter, is given by

$$\begin{gathered} H_0 : \text{The process is both a capable and predictable ARMA}(1,1)\text{ process} \\ \text{versus} \\ H_A : \text{The process is either not capable or not predictable.} \end{gathered} \tag{185}$$

A test statistic related to both capability and predictability will be used to test the capability hypothesis. We will proceed by identifying a sub-hypothesis for capability and predictability separately, and then combine the two sub-hypotheses into a single hypothesis.

Capability is routinely expressed in terms of a capability index. Without loss of generality, we assume that the true process capability can be expressed by the capability index, C, with the property that larger values of C imply greater capability. We further assume that the minimum acceptable process capability can be expressed as C_0. Then, the sub-hypothesis that the process is capable can be expressed as

$$\begin{gathered} H_0 : C \geq C_0 \\ \text{versus} \\ H_A : C < C_0. \end{gathered} \tag{186}$$

In the previous chapter, we showed that the true process capability, C, can be estimated. We will refer to the estimated capability as $\hat{C}$. The sub-hypothesis that the process is capable can be tested at the α_c level of significance by comparing $\hat{C}$ with an appropriately selected critical value, C_{crit}. If $\hat{C}$ is found to be less than C_{crit}, then there is sufficient statistical evidence at the α_c level of significance to refute the sub-hypothesis and conclude that the process is not capable. In general, a one-to-one relationship exists between the level of significance and critical values.

Unlike capability, predictability cannot be captured as easily in a statistic. However, for the ARMA(1,1) process we are interested in, the underlying errors of an adequately specified and estimated model are normally distributed, and we can test that normality in order to test the predictability of the process. In order to do so, suppose the underlying errors, ϵ_t, are independent and identically distributed according to some unknown distribution, F. Further suppose a measure of normality is given as W and its estimate is given as $\hat{W}$. The predictability sub-hypothesis can be tested at the α_p level of significance by comparing $\hat{W}$ with an appropriately selected critical value, W_{crit}. If $\hat{W}$ is less than W_{crit}, then there is sufficient statistical evidence at the α_p level of significance to refute the sub-hypothesis of normality and conclude that the process is not predictable. That is, the sub-hypothesis that the process is predictable can be expressed as

$$H_0 : W \geq W_0$$

versus

$$H_A : W < W_0. \tag{187}$$

The two null sub-hypotheses for capability and predictability can be easily combined into a bivariate hypothesis. A test space, $\mathcal{P}$, can be defined as the set of ordered pairs (c, w) such that c is an element in the range of C and w is an element in the range of W. The true capability and predictability of the process is given by $P = (C, W)$. The acceptable range for the test space corresponding to the null hypothesis, $P_0 = \{(c_0, w_0)\}$, is defined as the subset of $\mathcal{P}$ such that $c_0 \geq C_0$ and $w_0 \geq W_0$. Then, the combined capability hypothesis can be stated as

$$H_0 : P \in P_0$$

versus

$$H_A : P \notin P_0. \tag{188}$$

The test statistic, $\hat{P}$ is defined as the ordered pair $(\hat{C}, \hat{W})$. The critical region for the statistic is the subset of $\mathcal{P}$ for which either $\hat{C} < C_{crit}$ or $\hat{W} < W_{crit}$, or both.

Although critical values, C_{crit} and W_{crit}, corresponding to a specific levels of significance, α_c and α_p, can be selected for $\hat{C}$ and $\hat{W}$, determining the overall level of significance for the bivariate test is a complex task. If the individual tests were independent, then the overall level of significance would be given as $1-(1-\alpha_c)(1-\alpha_p)$. However, the predictability and capability of the process are generally related, and therefore, the individual tests are most likely dependent. In addition, if the test is applied sequentially to an autocorrelated process, the levels of significance corresponding to fixed critical values may change over time.

5.4 A Practical Method for Monitoring Process Capability

In this section, we propose a specific capability monitoring system which expands upon the general method introduced earlier. The proposed capability monitoring system depicted in Figure 13 incorporates specific design considerations into the general method shown in Figure 12. A more detailed description of the proposed system follows.

5.4.1 Parameter Definition. The parameters which must be explicitly specified are listed in the top block of Figure 13. These include upper and lower specification limits as well as a target value for the process. It is assumed that a measurable process exists. In addition, critical values for the statistical test of the capability hypothesis are assumed to be specified. Selecting appropriate critical values is not a trivial task and will be discussed in Section 5.4.4.

5.4.2 Process Start-up and Sampling Strategy. The next two blocks in Figure 13 capture the essence of the sampling strategy. In the proposed method, a moving window of the 30 most recent observations is maintained as the basis for computing the test statistics. At every iteration, the oldest observation in the window is dropped and the time-series of observations in the window is augmented by the current observation. The number of observations in the window is set at 30 to balance two conflicting objectives. First, a sufficient number of observations is required to adequately fit an ARMA(1,1) model. Here, an adequate fit is loosely defined as a fit which yields the ability to assess process capability. That is, we are not trying to determine the true parameters of the model so much as we are gathering information suitable for computing a capability statistic. This is much different than Box and Jenkins (1976) goal of determining the correct model from the family of all ARIMA models. For their much more difficult problem, they recommend a sample size of

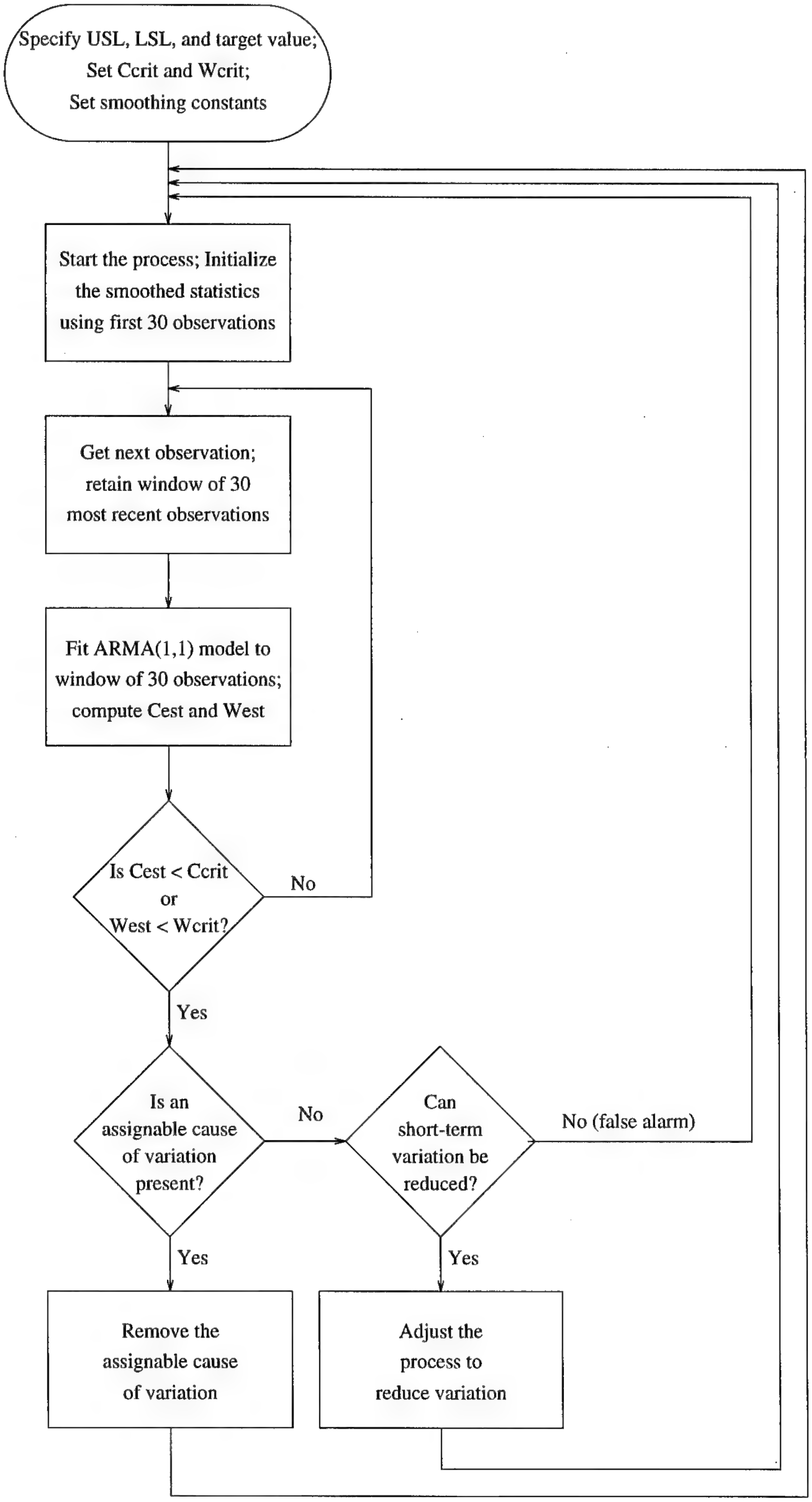

Specify USL, LSL, and target value;
Set Ccrit and Wcrit;
Set smoothing constants
Start the process; Initialize the smoothed statistics using first 30 observations
Get next observation; retain window of 30 most recent observations
Fit ARMA(1,1) model to window of 30 observations; compute Cest and West
Is Cest < Ccrit or West < Wcrit?
No
Yes
Is an assignable cause of variation present?
No
Can short-term variation be reduced?
No (false alarm)
Yes
Yes
Remove the assignable cause of variation
Adjust the process to reduce variation

Figure 13. Flowchart of a proposed method for monitoring process capability.

at least 50 observations. In our experience, a sample of 30 observations provides estimates for ϕ and θ which are almost as good as estimates based on 50 observations. Second, a small window size is desired to allow the method to respond more quickly to the introduction of an assignable cause. A large window size can be expected to hide to initial effects of an assignable cause until a sufficient number of observations affected by that cause are included in the window.

5.4.3 Compute the Capability Test Statistic. As mentioned for the general capability monitoring system, the capability test statistic fundamentally determines what is being monitored by the system. We develop the test statistic for the proposed system in this subsection. As a preliminary step, an ARMA(1,1) model is fit to the moving window of observations. The fitted model is then used to construct the test statistic.

5.4.3.1 Fitting an ARMA(1,1) Model. The capability hypothesis asserts, in part, that the observations arise from a process that can be approximated reasonably well by an ARMA(1,1) process. In order to construct a test statistic, we assume that the capability hypothesis is true, and estimate the parameters of the assumed ARMA(1,1) model. Given a set of consecutive observations from a time-series, it is relatively straightforward to fit an ARMA(1,1) model to those observations. The fitted model will yield estimates of the autoregressive and moving average parameters of the time-series, $\hat{\phi}$ and $\hat{\theta}$ respectively, as well as a location parameter, $\hat{\xi}$. In addition, the residuals from the fitted model can be used to estimate the underlying error terms. However, some criteria must be chosen by which to fit the model.

A frequently used criteria by which to fit an ARMA(1,1) model to a time-series is to minimize the sum of the square residuals from the fitted model. For the experiments documented in this chapter, the Matlab routine 'armax' is used to fit an ARMA(1,1) model to the moving window of observations. The 'armax' routine robustly minimizes the quadratic prediction error using an iterative Gauss-Newton algorithm (Ljung, 1992) and provides estimates of ϕ and θ. The 'armax' routine also provides an estimate of the variance of the error terms. Since we are limiting ourselves to the family of stationary ARMA(1,1) processes, we know that the true value of the parameter ϕ has magnitude less than one. In the implementation of the proposed system, the estimate of ϕ is limited to the range

$$-.97 \leq \hat{\phi} \leq .97. \tag{189}$$

This restriction is imposed to avoid problems encountered when the 'armax' routine attempts to fit parameters on the boundary of the feasible range for the parameters. When $\hat{\phi}$ has a magnitude greater than one, the fitted ARMA(1,1) process is not stationary and thus cannot be said to have an unconditional variance. When $\hat{\phi}$ is very close to one, the estimated unconditional variance becomes excessively large, and, for the proposed method results in false alarms.

The fitted model can be directly used to create a test statistic. However, since the fitted model is based upon a relatively small number of observations, the estimates for any given parameter tend to exhibit large variances. If the test statistic is directly calculated from the fitted model, it will also tend to have an unacceptably large variance and so lower critical values for the test statistic will have to chosen. This, in turn, will lead to less power when the test statistic is impacted by the occurrence of an assignable cause. The proposed system addresses this concern by smoothing the fitted parameters prior to computing the test statistic. For a given window size, the test statistic calculated from the smoothed fitted parameters exhibits less variance than does the test statistic from the unsmoothed fitted parameters. In our experience with ARMA(1,1) processes, the variance of the smoothed test statistic based upon a window of 30 observations is equivalent to the variance of the unsmoothed test statistic based upon fitting the ARMA(1,1) model with 50 observations. The smoothed parameter estimates, denoted $\tilde{\phi}$, $\tilde{\theta}$, $\tilde{\xi}$ and $\tilde{\sigma}_\epsilon$, of the true process parameters, denoted ϕ, θ, ξ and σ_ϵ, are computed as the exponentially weighted moving average of the fitted parameters, denoted $\hat{\phi}$, $\hat{\theta}$, $\hat{\xi}$ and $\hat{\sigma_\epsilon}$, via

$$\begin{aligned} \tilde{\phi} &= \lambda\hat{\phi} + (1-\lambda)\tilde{\phi} \\ \tilde{\theta} &= \lambda\hat{\theta} + (1-\lambda)\tilde{\theta} \\ \tilde{\xi} &= \lambda\hat{\xi} + (1-\lambda)\tilde{\xi} \\ \tilde{\sigma_\epsilon} &= \lambda\hat{\sigma_\epsilon} + (1-\lambda)\tilde{\sigma_\epsilon} \end{aligned} \tag{190}$$

where λ is the smoothing constant. Montgomery (1991) recommends selecting λ from the interval $0.05 \leq 0.25$. For the proposed system, we chose $\lambda = 0.15$. The use of Equation 190 requires initial values to be specified for each smoothed parameter. Initial values can naively be set to zero. However, when the true mean of the fitted parameter is not zero, that choice causes a delay of several observations before the smoothed parameter approaches its non-zero mean. A potentially better alternative is to use the true non-zero mean value as the initial value. The obvious problem with this alternative is that the mean may not be known. However, the fitted parameter from the start-up period (i.e. from the first 30 observations) are available. The fitted parameters from the process start-up period are used in the proposed method to select initial smoothed parameters, via

$$\begin{aligned} \tilde{\phi} &= 0.8\;\hat{\phi} \\ \tilde{\theta} &= 0.9\;\hat{\theta} \\ \tilde{\xi} &= 0 \\ \tilde{\sigma_\epsilon} &= 0.9\hat{\sigma_\epsilon}. \end{aligned} \tag{191}$$

Note that $\tilde{\xi}$ is initially set to zero since we assume that the process is well centered at start-up. The other initial fitted parameters are chosen close to, but slightly less than, the fitted parameters from the first window of observations in order to approximate the mean of the parameter estimates while avoiding the potential for overly large initial values due to variability in the estimates. We biased the initial fitted parameters in order to lessen the possibility of a false alarm on the first few observations.

5.4.3.2 Definition of the Proposed Capability Test Statistic. The proposed capability test statistic has two components: an estimate of process capability and an estimate of the process predictability. Both of these components are supported by our choice to fit an ARMA(1,1) process with a window of 30 observations. There are several choices of capability indices available to estimate the true process capability. The two most popular of these are C_{pk} and C_{pm}. This research uses the C_{pk} index as the basis for its test statistic.

Another decision that is critical to the development of the proposed system is selecting the time frame. In Chapter IV, we described time-specific, time-average and long-term capability. Since we don't know which future times might be of interest to the owner of the process, we chose not to implement our system using time-average capability. We also did not select the time-specific capability at the next time step since it ignores the behavior of the process in the distant (and not so distant) future. Instead, we chose to use long-term capability as the basis for our capability test statistic. Recall that the long-term analogue to the C_{pk} index is given by

$$C_x = \min\left[\frac{USL - \mu_x}{3\sigma_x}, \frac{\mu_x - LSL}{3\sigma_x}\right] \tag{192}$$

where μ_x is the long-term mean of the process and σ_x^2 is the long-term variance of the process.

Under the null hypothesis that the observations arise from an ARMA(1,1) process, we know that the unconditional process variance, σ_x^2, is related to the variance of the error terms, σ_ϵ^2, via

$$\sigma_x = \frac{1 + \theta^2 - 2\phi\theta}{1 - \phi^2}\sigma_\epsilon. \tag{193}$$

Using the results from fitting the ARMA(1,1) model, we can estimate long-term process variation via

$$\hat{\sigma}_x = \frac{1 + \tilde{\theta}^2 - 2\tilde{\phi}\tilde{\theta}}{1 - \tilde{\phi}^2}\tilde{\sigma}_\epsilon. \tag{194}$$

Similarly, an estimate of the long-term mean of an ARMA(1,1) process is given by

$$\hat{\mu} = \frac{\tilde{\xi}}{1 - \tilde{\phi}}. \tag{195}$$

Finally, long-term process capability is estimated as

$$\hat{\mathcal{C}}_{pk} = \min\left[\frac{USL - \hat{\mu}}{3\hat{\sigma}_x}, \frac{\hat{\mu} - LSL}{3\hat{\sigma}_x}\right]. \tag{196}$$

The normality of the errors cannot be directly tested since the error terms are unknown. However, the normality of the residuals from the fitted model can be tested. As we discussed earlier, we are not interested in normality, per se. Instead, we are interested in whether or not we have statistical evidence to indicate the appropriateness of predicting the future based upon the fitted model. If the residuals from the fitted model are (approximately) normally distributed, the statistical evidence tends to indicate that we can use the model for predicting. If they are not, then we have evidence suggesting it is not appropriate to predict the future using the model. The Shapiro-Wilk test for normality is a very general statistical test for testing the normality of a distribution with an unspecified mean and variance (Conover, 1980; Shapiro, 1980). In some instances, an assignable cause of variation will be reflected as one or more outliers in the stream of errors. Igelwicz and Hoaglin (1993) present the Shapiro-Wilk test as one method for identifying such outliers.

The Shapiro-Wilk test statistic is suitable for small samples; tables of critical values for samples of up to 50 observations are available (Conover, 1980). Given a random sample $\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$ of size n with some unknown distribution function $F(\epsilon)$, the Shapiro-Wilk test for normality tests the null hypothesis that $F(\epsilon)$ is a normal distribution function with unspecified mean and variance against the alternative hypothesis the $F(\epsilon)$ is non-normal. In order to compute the Shapiro-Wilk test statistic, first, let $\bar{\mathcal{E}}$ be the sample mean and let $\mathcal{E}_{(i)}$ denote the ith order statistic so that

$$\mathcal{E}_{(1)} \leq \mathcal{E}_{(2)} \leq \ldots \leq \mathcal{E}_{(n)}. \tag{197}$$

Then, the denominator D of the Shapiro-Wilk test statistic is given via

$$D = \sum_{i=1}^{n} (\mathcal{E}_i - \bar{\mathcal{E}})^2 \tag{198}$$

and the test statistic T is given by

$$T = \frac{1}{D} \left[\sum_{i=1}^{k} a_i (\mathcal{E}_{(n-i+1)} - \mathcal{E}_{(i)}) \right]^2 \tag{199}$$

where k is approximately $n/2$ and the coefficients $a_1, a_2, \ldots, a_k$ are provided (e.g. Table A17 in Conover). The Shapiro-Wilk statistic can be converted to one that has an approximate normal distribution via the transformation

$$G = b_n + c_n \ln\{(T - d_n)/(1 - T)\} \tag{200}$$

where b_n, c_n and d_n are provided (e.g. Table A19 in Conover). In this form, it is straightforward to select a critical value based upon any specified level of significance using the properties of the normal distribution. For example, the critical value for G corresponding to an α level of significance is $\Phi^{-1}(\alpha)$, where Φ^{-1} is the inverse cumulative distribution function of the standard normal distribution.

Since the true underlying error terms are unknown, the Shapiro-Wilk test statistic, T, cannot be directly determined. Instead, the statistic will be computed as a function of the fitted residuals and will be denoted $\hat{W}$. For the proposed system,

$$\hat{W} = \Phi(G) \tag{201}$$

where G is computed using the residuals of the fitted model.

Although we assume that the errors driving the ARMA(1,1) model are independent and normally distributed, we generally cannot make that assumption about the residuals from the fitted model. Since the model is fit so as to minimize the sum of the square residuals, the fitted residuals may not be normally distributed. However, as discussed in Chapter 4, when enough observations are available and the errors are normally distributed, the residuals from a properly specified model will approach the true errors, and so will approach a normal distribution. For the proposed system, we make the assumption that the residuals from the fitted model do approach a normal distribution in the absence of any assignable causes of variation. An assignable cause of variation may have (but is not limited to) two effects on the residuals. First, the assignable cause may be reflected by a small, persistent change in the residuals. In this case, while the time at which the assignable cause occurred is within the moving window, the residuals may not be (approximately) normally distributed. After the moving window has completely passed the time of the assignable cause, the residuals may return to a normal distribution and so would not be detected by the Shapiro-Wilk test. However, this shift should be reflected in a change in the estimate of the variance of the errors, and thus in the estimate of capability. Although the change in estimated capability may not result in an immediate signal indicating the presence of an assignable cause, it should increase the chances of such a signal in the future. The second way an assignable cause may be reflected in the residuals is by a (hopefully) large increase in one or more of the residuals. The large change in one (or perhaps a few) residuals is the case we are hoping to detect by using the Shapiro-Wilk statistic. This type of large change in the residuals is also what the special cause chart is designed to detect.

5.4.4 Comparing the Test Statistic to its Critical Region. In order to compare the results from our proposed system to other methods that attempt to monitor autocorrelated processes, and in particular ARMA(1,1) processes, we attempted to achieve an average run length of at least 370 in the absence of any assignable causes of variation over the range of ARMA(1,1) models. Achieving such average run lengths required a balance between the critical regions for the test statistics $\hat{C}$ and $\hat{W}$. To maintain a constant average run length in the absence of assignable cause variation, any increase in one critical value must be offset by a decrease in the other critical value. Our original desire was to achieve some parity between the number of false alarms attributable to each portion of the test. However, through trial and error, we learned that it would not be possible to achieve that parity. A Shapiro-Wilk test applied to an independent normal set of observations with a critical value of $W_{crit} = 1/370$ will, by definition, result in a false alarm approximately 1 out of 370 times. This critical value was selected for its intuitive value, although, when applied to residuals it generally would produce fewer false alarms. The critical value, $\hat{C}$, was then chosen to give an average run length for the combined test of at least approximately 370 throughout our design region. The method proposed in this chapter uses $C_{crit} = 0.90$ and $W_{crit} = 1/370$.

5.5 Results.

The capability monitoring system proposed in this chapter was tested by monitoring simulated ARMA(1,1) time-series subjected to a variety of assignable causes. A copy of the Matlab code used to both implement the system and conduct the simulation runs is contained in Appendix D. At the start of this section, details of the simulation are presented and are followed by analysis and interpretation of the results obtained for a variety of assignable causes.

5.5.1 Details of the Simulation. In order to test the proposed capability monitoring system, we ran a series of simulated experiments. Each design point for the experiment consisted of a combination of the autoregressive and moving average parameters for the process truth model and an assignable cause of variation. For ease of comparison to results from other approaches to monitoring ARMA(1,1) models, we selected the design points previously used by Wardell, Moskowitz and Plante (1990, 1994). The design points span the family of stationary ARMA(1,1) models. Experimental settings for ϕ are from the set $\{-.95,\ -.475,\ 0,\ .475,\ .95\}$ and settings for θ are from the set $\{-.90,\ -.45,\ 0,\ .45,\ .90\}$. We conducted one thousand runs at each design point to estimate an average run length. This number of runs proved sufficient for the analysis that follows, giving 90 percent confidence intervals of approximately ± 10 percent of the average run length in the absence of assignable causes and much tighter intervals in the presence of large assignable causes.

Each run began by randomly selecting an observation using the unconditional density function of the truth model. Additional observations were generated by applying the truth model using simulated standard normal errors. This procedure was repeated until a series of 30 observations were generated within the specification limits. In the event one or more of the initial 30 observations occurred outside of the specification limits, the observations were discarded and the run was restarted.

Once a initially controlled time-series was generated, the applicable assignable cause was imposed upon subsequent observations. Additional observations were generated until the capability monitoring system signaled a lack of capability. The number of additional observations was recorded as the run length for the run. The average run length was computed as the average of the run lengths at each design point. A 90 percent confidence interval for the mean run length is given by

$$ARL \ \pm \frac{t(.05, n-1)SRL}{\sqrt{n}} \qquad (202)$$

where ARL is the average run length, SRL is the sample standard deviation of the run lengths, n is the number of runs, and $t(.05, n-1)$ is the $\alpha = .05$ value of the t distribution with $v = n - 1$ degrees of freedom.

5.5.2 Average Run Length Results when No Assignable Cause is Present. In any monitoring system, we would like the average run length in the absence of any assignable cause to be large while the average run length in the presence of the an assignable cause to be small. In practice, an average run length of 370 observations in the absence of any assignable cause is often used.

Table 13 contains results for the proposed system in the absence of assignable causes of variation. Each average run length presented in the table is derived from one thousand simulated runs. Since no assignable cause of variation is present, each run is stopped when a false alarm is signaled. The false alarm can arise from either a low estimate of process capability or an indication that the fitted model is not suitable for prediction. The percent of false alarms arising from each part of the capability test is listed in the table.

In Table 13, we can see that the average run length exceeds the desired minimum of 370 in 23 of the design points. In 2 design points, the average run length is slightly below 370 (i.e. 366.9 and 361.0), however, the 90 percent confidence interval for the ARL includes 370 at every design point. The capability portion of the test accounts for approximately half for the false alarms at several design points, although, at most design points, the majority of the false alarms are due to the predictability portion of the test.

5.5.3 Average Run Length Results in the Presence of an Assignable Cause. While it is important to provide a dependable minimum average run length in the absence of assignable causes of variation, the ultimate test of the proposed method is its ability to react to the introduction of an assignable cause of variation. A standard assignable cause prevalent in the literature is a shift in the mean of the process. The magnitude of the mean shift is generally measured in multiples of the standard deviation of the process. Table 14 contains the results derived by simulating two hundred runs for the proposed system in the presence mean shifts ranging in size from one-half standard deviation to three standard deviations.

Table 13. Simulation results for ARL in absence of any assignable causes of variation.

		$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$	ARL	688.0	747.7	695.2	481.2	610.2
0.90	FAP	98.0	97.0	88.5	63.5	75.0
	LO 90	620.2	663.7	625.2	427.5	537.9
	HI 90	755.8	831.7	765.2	534.9	682.5
$\theta =$	ARL	517.3	704.2	714.3	538.6	515.8
0.45	FAP	75.5	99.0	95.0	70.0	61.5
	LO 90	463.0	634.2	637.6	485.1	459.0
	HI 90	571.7	774.2	790.9	592.1	572.6
$\theta =$	ARL	539.1	623.4	716.9	684.5	380.7
0.00	FAP	67.0	78.0	99.0	82.0	49.5
	LO 90	482.1	549.4	644.5	613.9	336.2
	HI 90	596.2	697.4	789.3	755.1	425.2
$\theta =$	ARL	499.2	471.0	684.6	695.6	366.9
-0.45	FAP	70.5	59.5	87.5	99.5	53.0
	LO 90	433.3	412.2	613.0	631.4	326.5
	HI 90	565.0	529.8	756.3	759.9	407.3
$\theta =$	ARL	640.8	361.0	586.9	717.0	719.0
0.90	FAP	65.5	47.0	73.0	95.0	100.0
	LO 90	562.7	316.2	523.3	636.0	646.0
	HI 90	719.0	405.9	650.5	798.1	792.0

ARL is the average run length from one thousand runs.
FAP is the percent of false alarms due to the predictability portion of the capability test.
LO 90 is the lower limit of a 90 percent confidence interval on ARL.
HI 90 is the upper limit of a 90 percent confidence interval on ARL.

A direct comparison to the documented results for other monitoring methods is not appropriate since those other results generally assume that the parameters of the ARMA(1,1) model are known. For instance, the results published by Wardell, Moskowitz and Plante (1994) were developed with full knowledge of the truth model for every design point. A copy of their key results, documenting the average run length after various mean shifts for the special cause, X, and EWMA control charts is contained in Appendix E. They acknowledge that "the limits of the Shewhart and EWMA charts had to be modified, sometimes substantially, to obtain an in-control ARL of about 370 for each (design point)." In effect, we are comparing our single proposed system against twenty-five different versions of the special cause chart, X chart and EWMA chart each incorporating additional information about the parameters of the truth model that our proposed system does not have available. In the next section, we examine the added value of perfect model information.

General trends, common to the results documented in both Table 14 and Appendix E, can be identified. However, the reader is cautioned that point-by-point comparison is akin to comparing apples and oranges. While the proposed capability monitoring system does not use knowledge about the parameters of the truth model, we contend that it provides a predictable response over the entire range of stationary ARMA(1,1) models which is, in a broad sense, similar to that of other charts that do use the additional knowledge. That is, in the absence of assignable causes of variation, the average run lengths for all of the compared techniques are at least 370, while the introduction of a significant assignable cause (e.g. a shift in the mean) results in average run lengths that are significantly less than 370.

5.6 The Value of Knowing the Parameters of the Truth Model.

In this section, we examine the value of perfect model information to a capability monitoring system. There are two major benefits to knowing the truth model. First, the error terms do not have to be estimated, instead, exact residual values can be determined based upon the model and the observations. Second, critical values can be selected to ensure a desired average run length in the absence of assignable cause variation. In addition, if a particular assignable cause can be anticipated, then the statistical test can potentially be optimized to respond to that exact assignable cause. In effect, a control chart is an optimized version of a general capability monitoring method.

5.6.1 Comparing Control Charts to a General Capability Monitoring Method. In a very real sense, the general capability monitoring method can be considered a superset

Table 14. Average Run Length from simulation for mean shifts of various multiples of σ_x.

	$\Delta\sigma_X$	$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.90	0.0	688.0	747.7	695.2	481.2	610.2
	0.5	524.9	435.6	314.2	188.7	72.4
	1.0	150.7	99.6	78.0	61.0	1.7
	1.5	43.6	30.2	31.1	23.7	1.1
	2.0	20.2	18.1	17.5	10.8	1.0
	2.5	15.1	14.1	12.6	5.1	1.0
	3.0	11.8	11.3	9.9	3.1	1.0
	$\Delta\sigma_X$	$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.45	0.0	517.3	704.2	714.3	538.6	515.8
	0.5	432.0	542.3	400.5	234.3	103.2
	1.0	226.6	126.8	95.6	77.5	2.6
	1.5	113.0	35.8	32.2	25.8	1.3
	2.0	47.2	19.2	17.8	12.9	1.0
	2.5	15.7	15.0	14.4	9.1	1.0
	3.0	5.6	11.3	10.8	4.5	1.0
	$\Delta\sigma_X$	$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.00	0.0	539.1	623.4	716.9	684.5	380.7
	0.5	459.5	337.5	537.8	352.0	201.7
	1.0	206.8	109.2	128.9	100.0	28.0
	1.5	42.9	39.0	36.8	32.3	2.7
	2.0	4.0	18.4	19.6	18.8	1.2
	2.5	2.1	13.3	14.7	13.4	1.0
	3.0	1.0	10.3	11.1	9.4	1.0
	$\Delta\sigma_X$	$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ -0.45	0.0	499.2	471.0	684.6	695.6	366.9
	0.5	371.3	246.8	358.6	517.2	247.0
	1.0	104.2	95.1	99.2	125.5	125.4
	1.5	3.9	38.2	35.2	35.2	49.9
	2.0	1.0	17.7	18.4	19.2	14.9
	2.5	1.0	11.1	13.4	14.2	3.8
	3.0	1.0	7.4	10.6	11.6	1.6
	$\Delta\sigma_X$	$\phi = 0.95$	$\phi = 0.475$	$\phi = 0.0$	$\phi = -0.475$	$\phi = -0.95$
$\theta =$ 0.90	0.0	640.8	361.0	586.9	717.0	719.0
	0.5	409.7	196.5	292.3	407.3	525.7
	1.0	45.6	79.1	78.1	96.6	140.3
	1.5	1.0	33.7	30.2	32.7	36.3
	2.0	1.0	16.1	15.6	17.3	18.8
	2.5	1.0	7.7	10.9	12.5	14.4
	3.0	1.0	4.3	8.6	10.4	12.0

which spans all of the standard control charts. Our intuitive reliance on capability indices helps to explain this relationship by expressing capability as a function of both process spread and process location. Control charts, on the other hand, monitor the state of statistical control by monitoring either process spread or process location. Consider the three control charts examined by Wardell, Moskowitz and Plante (1994):

- *The Special Cause Chart.* The special cause chart tracks the residuals of a known model. It signals an alarm when the magnitude of the current residual exceeds three times the variance of the underlying errors. In effect, the special cause chart monitors the one step ahead variance while assuming that the one step ahead expected value is equal to the target value. Thus, the special cause chart can be considered a simple instantiation of a one step ahead capability monitoring system.
- *The X-Chart.* The X-Chart tracks the actual process observations. It signals an alarm when the current observation falls outside of predetermined control limits. The current observation can be considered to be a naive estimate of the process mean. A capability monitoring system equivalent to the X-chart can be identified in which the current process observation is used as the one step ahead expected value of the process and the one step ahead variance of the process is fixed. Thus, the X-chart can be considered to be an instantiation of a one step ahead capability monitoring system.
- *The EWMA Chart.* Like the X-chart, the EWMA chart tracks an estimate of the process location. By incorporating information from (all) previous observations, the EWMA is intended to provide a more robust estimate of the process location. At its heart, however, the EWMA is still just an estimate of the one step ahead expected value of the process. Like the X-chart, the process variance is assumed to be fixed and the EWMA chart can be considered to be an instantiation of a one step ahead capability monitoring system.

The implied benefit of the capability monitoring approach is that monitoring both process location and process spread should provide more robust abilities to detect a variety of assignable causes.

5.6.2 Capability Monitoring Applied to an Independent Normal Process. The X-chart and EWMA chart were originally created to detect changes from an independent normal process. To demonstrate the value of additional information about the truth model, a modified capability monitoring method will be compared with those charts for both shifts in the process mean and process variance. Note that the special cause chart is equivalent to the X-chart under the assumptions of independence and normality. The modifications to the proposed monitoring system are three-fold. First, the estimate of the process variance is computed using the adjusted mean absolute deviation, or MAD, of the moving window of n underlying error terms, denoted ϵ_1 to ϵ_n, via

$$MAD = 1.25\frac{1}{n}\sum_{i=1}^{n}|\epsilon_i|. \tag{203}$$

The adjusted MAD is an unbiased estimator of the true standard deviation of the errors when the errors are independent and normally distributed (Montgomery et al., 1990). Second, since model fitting is not required, there is also no need to smooth model parameters. Finally, the critical region for the capability test was changed to give an in-control average run length of approximately 370. The critical values used for this experiment, $C_{crit} = .96$ and $W_{crit} = 1/400$, were chosen by trial and error.

The average run length from one thousand simulated runs of the modified capability monitoring method for a variety of shifts in the mean and shifts in the process variance are contained in Tables 15 and 16. Table 15 includes results for the special cause chart, X-chart and the EWMA chart from Appendix E. Note that, for the case of independent and identically distributed observations, the special cause chart is identical to the X-chart. Table 16 includes theoretic results for the special cause and X-charts with 3σ control limits and simulated results from 10000 runs of an EWMA chart with a critical value of .622. The same information is graphically depicted in Figure 14. In the figure, the average run lengths are normalized by dividing the tabulated average run length for each combination of method and assignable cause by the average run length for the method that detects the assignable cause fastest. For example, a value of one indicates that the method is the best at detecting that assignable cause, while a value of two indicates that the method takes, on average, twice as long to detect that assignable cause as the best method.

Not surprisingly, the EWMA is clearly superior at doing what it was designed to do: detect shifts of $2\sigma_x$ or less in the process mean. The special cause and X-charts provide a shorter average run length than the EWMA for larger shifts in the process mean. Furthermore, the special cause and X-charts provide a shorter average run length than the EWMA for any size increase in the process standard deviation. Like the EWMA, the modified capability method outperforms the special cause and X-charts for mean shifts of $2\sigma_x$ or less, but does not equal the EWMA's performance in that region. However, the modified capability method is superior to the EWMA for any size change in the process variance. It also outperforms the special cause and X-charts for increases of 50 percent or less in the process standard deviation. In total, the modified capability monitoring attempts to strike a happy medium, being able to detect a variety of shifts in either the process location or the process spread.

5.6.3 Capability Monitoring Applied to a Known ARMA(1,1) Process. When the parameters of the truth model which generates the process observations in the absence of assignable cause variation is known, a monitoring system can be tailored to that truth model. For example, the ARMA(1,1) model can be defined by the parameters ϕ, θ, ξ and σ_ϵ. Given an initial observation and its associated error, the residuals from all future observations can be directly calculated from equation 57. Since the special cause chart monitors the residuals from a process, all of this information is required in order to apply it to an ARMA(1,1) process. While the parameters and initial underlying errors are not

Table 15. Simulation and previously published results for ARL given independent normal observations for a variety of mean shifts using process knowledge.

Mean Shift in	Method			
Multiples of σ_x	Capability[a]	SCC[b]	X[b]	EWMA[b]
0.0	380.3	370.4	370.4	369.0
0.5	48.8	155.2	155.2	28.2
1.0	14.5	43.9	43.9	9.7
1.5	8.4	14.9	14.9	5.8
2.0	5.8	6.3	6.3	4.2
2.5	4.4	3.2	3.2	3.3
3.0	3.3	2.0	2.0	2.8

[a] Listed ARL is from 1000 simulated runs.

[b] Listed ARL previously published (see Appendix E).

Table 16. Simulation results for ARL given independent normal observations for increases in the standard deviation of the process error.

Increase in	Method			
Std Dev (%)	Capability	SCCC	X	EWMA
0	380.3	370.4	370.4	370.6
25	41.0	61.0	61.0	99.8
50	19.4	22.0	22.0	46.8
100	9.7	7.5	7.5	19.3
200	4.9	3.2	3.2	8.0
300	3.2	2.2	2.2	4.9

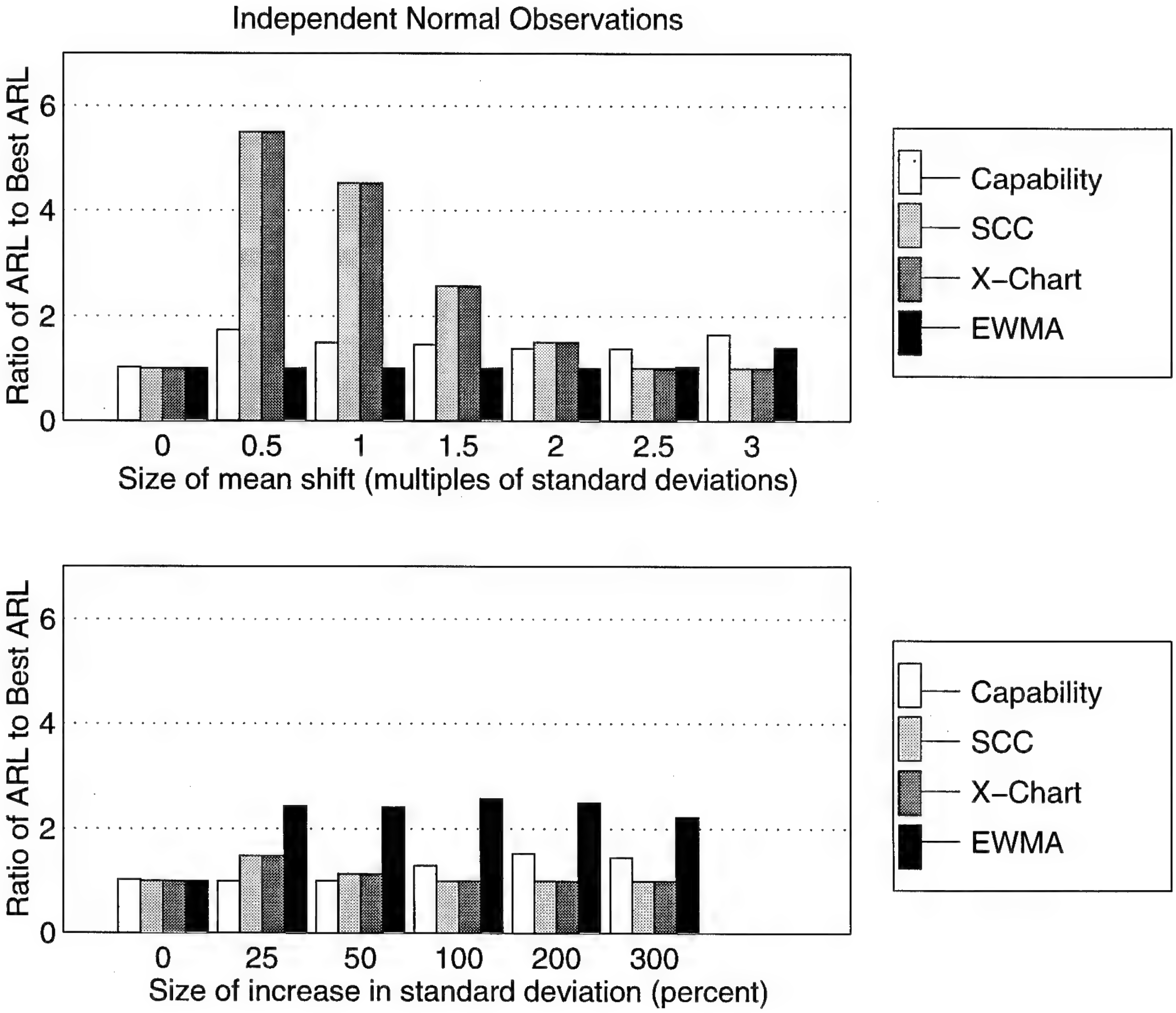


Figure 14. Comparison between capability monitoring system, Special Cause Chart, X-Chart and EWMA Chart for independent normal observations.

generally known, in practice they are estimated by fitting an ARMA(1,1) model to historical data. In this section, we develop a capability monitoring system for an ARMA(1,1) process with known parameters. The average run length of that system after a variety of assignable causes is compared to the average run lengths from the special cause, EWMA and X-charts.

In order to effectively use the additional knowledge about the model parameters and to keep in the spirit of the special cause, EWMA and X-charts, the capability monitoring system in this section will use one-step ahead capability rather than long-term capability. An estimate of the one-step ahead capability, $\hat{\mathcal{C}}_{T+1|T}$, requires an estimate of the one-step ahead expected value and one-step ahead variance. Given the known parameters, the current observation, and the current error (residual), the conditional one step ahead expected value is easily calculated. For the general capability monitoring system presented earlier in this chapter, the residuals from a moving window of 30 observations are used to estimate the one-step ahead variance. As the window passes over a given observation, 30 different residuals are generated for that time. When the model parameters are known, a single residual is determined for each point in time. In this case, an exponentially weighted moving variance can be used to estimate the one-step ahead variance. The exponentially weighted moving variance at time t, denoted s_t^2, is given by

$$s_t^2 = \lambda r_t^2 + (1-\lambda)s_{t-1}^2 \tag{204}$$

where λ is a smoothing constant generally chosen between 0.05 and 0.3 and r_t is the residual at time t (MacGregor and Harris, 1993). The one-step ahead standard deviation can then be estimated by

$$\hat{\sigma}_{\epsilon,T+1|T} = s_T. \tag{205}$$

The exponentially weighted moving variance provides two benefits over the mean absolute deviation. First, since recent residuals are weighed more heavily than past residuals, a change in the residuals due to an assignable cause may be reflected in the estimate of the standard deviation of the errors more quickly than waiting for the moving window to accumulate enough of the affected residuals. Second, a single large residual due to the introduction of an assignable cause may increase the estimate of the error sufficiently to signal the assignable cause immediately. Thus, the predictability portion of the capability hypothesis may not need to be tested separately. A single test statistic for this case is

$$\hat{C}_{pk,T+1|T} = \frac{\min(USL - \hat{\mu}_{T+1|T},\ \hat{\mu}_{T+1|T} - LSL)}{3s_T}. \tag{206}$$

We tested the modified capability monitoring system for an ARMA(1,1) process with $\phi = .95$ and $\theta = .45$ (and, without any loss of generality, $\xi = 0$ and $\sigma_\epsilon = 1$). The critical value and smoothing constant used for this experiment, $C_{crit} = .85$ and $\lambda = 0.3$, were chosen by trial and error in order to achieve an average run length of approximately 370 in the absence of any assignable causes. Tables 17 and 18 depict the results from one thousand runs of the modified capability monitoring system and the results published for other methods. The same information is graphically depicted in Figure 15. As Figure 15 clearly shows, the modified capability monitoring system has the potential to respond rapidly to a variety of assignable causes. In particular, when compared to the special cause, EWMA and X-charts, the modified capability monitoring system has a lowest average run length for shifts in the mean of between $1.5\sigma_x$ and $2\sigma_x$. More significantly, the average run length from the modified capability monitoring system is within 25 percent of the best average run length from any of the three other methods. Furthermore, the modified method clearly outperforms the special cause chart for shifts in the mean of between $1.5\sigma_x$ and $2\sigma_x$, outperforms the X-chart for any size increase in error variance or shifts in the mean larger than $1.5\sigma_x$, and outperforms the EWMA for shifts in the mean larger than $1.5\sigma_x$. While the modified method is not superior to every other standard method for every assignable cause, it may be fair to say that, when compared to any one of the other methods, it is competitive for every type assignable cause and is superior for at least some types of assignable causes.

Table 17. Simulation and previously published results for ARL given observations from an ARMA(1,1) process with known parameters $\phi = .95$, $\theta = .45$, $\xi = 0$ and $\sigma_\epsilon = 1$ for a variety of mean shifts.

Mean Shift in	Method			
Multiples of σ_x	Capability[a]	SCC[b]	X[b]	EWMA[b]
0.0	367.8	370.4	392.9	362.9
0.5	253.0	349.7	262.7	232.2
1.0	122.9	274.7	108.7	93.8
1.5	46.5	147.6	50.9	45.6
2.0	14.1	43.5	20.8	23.6
2.5	3.9	6.6	7.3	13.5
3.0	1.5	1.3	2.2	8.8

[a] Listed ARL is from 1000 simulated runs.

[b] Listed ARL previously published (see Appendix E).

Table 18. Simulation results for ARL given observations from an ARMA(1,1) process with known parameters $\phi = .95$, $\theta = .45$, $\xi = 0$ and $\sigma_\epsilon = 1$ for a variety of increases in the standard deviation of the process error.

Increase in	Method			
Std Dev (%)	Capability[a]	SCC[b]	X[a]	EWMA[c]
0	367.8	382.0	370.4	N/A
25	74.3	61.0	107.1	N/A
50	28.3	22.0	49.7	N/A
100	10.3	7.5	20.2	N/A
200	4.5	3.2	7.2	N/A
300	3.0	2.2	4.2	N/A

[a] Listed ARL is from 1000 simulated runs.

[b] Theoretically derived results: ARL $= 1/(1 - \Phi(3/I) + \Phi(-3/I))$; $I = 1 + \%$ increase.

[c] Appropriate parameters for the EWMA chart are not available.

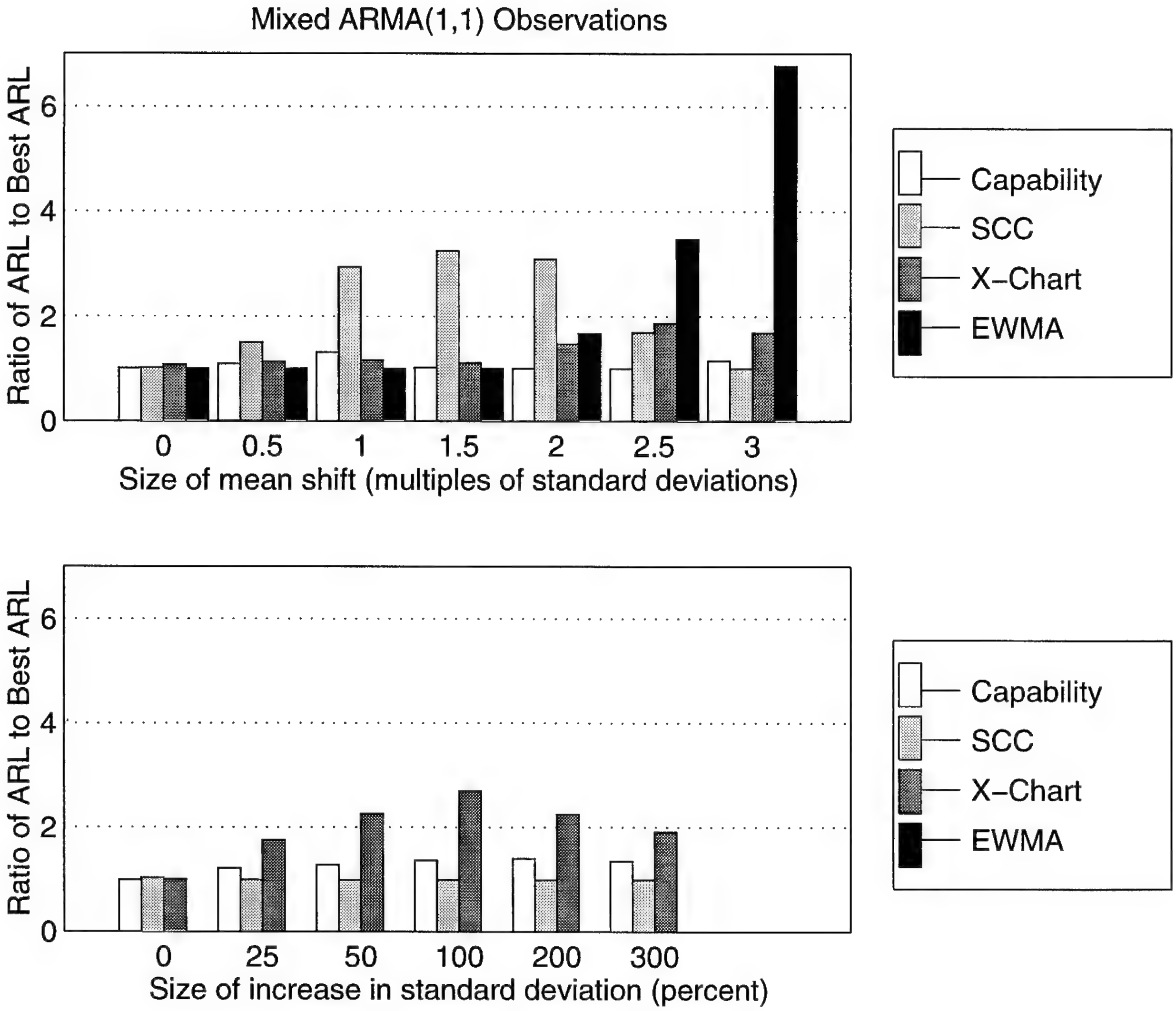


Figure 15. Comparison between capability monitoring system, Special Cause Chart, X-Chart and EWMA Chart for mixed ARMA(1,1) observations with $\phi = 0.95$ and $\theta = 0.45$.

5.7 Chapter Summary.

Capability monitoring is an evolutionary step in the field of statistical process control. In this chapter, we demonstrated the feasibility of monitoring process capability by comparing the results from a proposed capability monitoring method to published results for standard control charts. In general, capability monitoring methods differ from control chart methods by explicitly accounting for chance, assignable and structural causes of variation. In addition, capability monitoring methods directly monitor a measure of process quality rather than the state of statistical control. We showed that control charts are a subset of capability monitoring methods.

The proposed capability monitoring method achieved the goals we established for practical quality improvement tools. Over the range of stationary ARMA(1,1) models, it possesses an average run length of greater than 370 in the absence of assignable cause variation, while responding to changes in the process mean. Further, an initial implementation of the proposed system was able to process more than one observation per second on a Sparcstation 2, demonstrating the suitability of implementing the method on-line. An important advantage of the method over the special cause, X, and EWMA charts is that the parameters of the true ARMA(1,1) process do not need to be known prior to starting the system.

Even more compelling, when information about the parameters of the truth model is known, a capability based method for monitoring a process can respond as well as, and sometimes better than, existing methods based on statistical control. We demonstrated this comparison for both an independent normal process and a mixed ARMA(1,1) process. For these cases, a capability based monitoring system performed (almost) as well as the best of the standard control charts for any size mean shift or increase in error standard deviation. Furthermore, the capability system was clearly superior to each standard control chart for some assignable causes.

VI. Summary and Recommendations

6.1 Summary

The research documented in this dissertation has demonstrated the ability to monitor the capability of a process. The shifting emphasis from the 'state of statistical control' to 'capability' reflects a new paradigm. The need for this paradigm is evident in the documented limitations of control-based monitoring systems in the presence of autocorrelated process observations. The new paradigm fits within a larger philosophy of continuous process improvement based upon statistical thinking.

The standard control charts were developed based upon a two-source model of process variation. We propose a new taxonomy of three causes of variation. Our taxonomy explicitly accounts for structural cause variation, that is manifested by autocorrelation, as well as chance and assignable causes of variation. By breaking out variation in this way, a deeper understanding of the time-varying aspects of autocorrelated processes can be gained. By correctly analyzing the statistical properties of the process and taking the appropriate corrective actions, process variation can be reduced and product quality improved.

By accounting for structural cause variation, fixed control limits for processes from the family of stationary ARMA(1,1) models that achieve specified average run lengths in the absence of assignable cause variation can be determined. The control limits can be selected by incorporating conditional information gained about the process over time into the probability density function of the state of the process. The results of this part of the research can be used by quality practitioners to choose appropriate control limit multipliers for ARMA(1,1) processes.

A more significant aspect of this research is the development of a capability based monitoring method. The mathematical foundation for this method reveals that capability can be viewed as a time-varying attribute of a process. In this light, capability at some time in the future can be estimated using estimates of future process variation and location. Capability can be monitored by identifying the expected long-term capability of the process. Significant changes to the process will be reflected by a decrease in the estimated long-term process capability. A specific capability monitoring system is proposed which addresses the practical needs for quality improvement. The proposed system is shown to be comparable to a gamut of standard control charts over the family of ARMA(1,1) models. In addition, the utility of a capability based monitoring system that uses knowledge about the parameters of the truth model for the process is demonstrated for both an independent normal model and a mixed ARMA(1,1) model. The performance of the capability based systems are comparable to, and in some cases better than, the special cause chart, X-chart and EWMA chart.

6.2 Recommendations

The capability monitoring method proposed in this paper represents a first step in the new capability based paradigm. The opportunities for improving the method are numerous. Some opportunities include:

- Application of the method to real world data sets to demonstrate its suitability. The worth of any quality monitoring system is shown by its profitability.
- Conducting sensitivity analysis of the proposed system. For example, testing the system for processes other than those generated by the ARMA(1,1) model.
- Taking advantage of knowledge gained about the process over time to dynamically select the critical values for the capability test. Using additional knowledge might further enhance the detection capabilities of the system to allow direct comparison with documented results for other systems.

- Further development of the system under perfect process knowledge. This might include optimizing a capability monitoring system for a real world data set as well as conducting sensitivity analysis for a known model.
- Delivery of a packaged system able to be run in the field by a quality practitioner for real-world processing.

Appendix A. Summary of Notation used in Chapter III.

x_t	Observation at time t
X_t	Random variable for the observation at time t
ϵ_t	Underlying error at time t
$\mathcal{E}_t$	Random variable for the underlying error at time t
ϕ	Autoregressive parameter for an ARMA(1,1) model
θ	Moving average parameter for an ARMA(1,1) model
ξ	Location parameter for an ARMA(1,1) model
σ_x^2	Unconditional variance of the observations
σ_ϵ^2	Unconditional variance of the underlying errors
μ	Location parameter for the general linear filter
ψ	Coefficients for the general linear filter
$f(x_t)$	Probability density function of X_t
$f(x_t, \epsilon_t)$	JPDF of $(X_t, \mathcal{E}_t)$
$g(x_{t+1}, \epsilon_{t+1})$	JPDF of $(X_{t+1}, \mathcal{E}_{t+1})$
$f^*(x_t)$	Marginal PDF of X_t
$g^*(x_{t+1})$	Marginal PDF of X_{t+1}
$f_n(x_t, \epsilon_t)$	Conditional JPDF of $(X_t, \mathcal{E}_t)$ given that the n most recent observations have been within the specified control limits
$g_{n+1}(x_t, \epsilon_t)$	Conditional JPDF of $(X_t, \mathcal{E}_t)$ given that the n previous observations have been within the specified control limits
Φ	Cumulative density function of the error terms
y_1	Placeholding variable equivalent to x_{t+1}
y_2	Placeholding variable equivalent to ϵ_{t+1}
A_y	Region of integration

α	False alarm rate (for the iid case)
ARL	Average run length (for the iid case)
α_t	False alarm rate at time t
ARL_t	Average run length at time t
α_{30}	False alarm rate after at least 30 initially controlled observations
$ARL_{t\|30}$	Average run length at time t given at least 30 initially controlled observations
PDF	Probability density function
JPDF	Joint probability density function
iid	Independent and identically distributed

Appendix B. Matlab Code for Numerically Approximating the Joint Probability Density Function for an ARMA(1,1) Process.

```
start=clock;
phipoints = [0.95 0.475 0 -.475 -.95];
limpoints = [1.5 1.75 2 2.25 2.5 2.75 3 3.25];
phipoints = [0.95];
theta   = .9 ;
nrx     = 361;
nre     = 123;
nrloops = 30;

disp (sprintf ('nrx = %i  nre = %i  nrloops = %i',nrx, nre, nrloops));
nrphipoints = size(phipoints,1)*size(phipoints,2);
nrlimpoints = size(limpoints,1)*size(limpoints,2);
indexsize = nrphipoints*nrlimpoints;
arl = zeros (indexsize, nrloops+1);
for phiindex = 1:nrphipoints,
for limindex = 1:nrlimpoints,
index = (phiindex-1)*nrlimpoints + limindex;
phi = phipoints(phiindex);
lim = limpoints(limindex);
phicode(index) = phi;
limcode(index) = lim;

gam0 = (1 + theta^2  - 2*phi*theta) / (1 - phi^2);
sigmax = sqrt(gam0);
```

```
corr= 1/sigmax;
rho = (1-theta*phi)*(phi-theta)/(1+theta^2-2*phi*theta);
lcl = -lim*sigmax;
ucl =  lim*sigmax;
lcle = -5.5;
ucle = 5.5;
dopt = 1 - (normcdf(ucl) - normcdf(lcl));
d2 = dopt / (dopt+1);
tiny = d2 / (nrx * nre);
%  tiny = 0.0000000000001;

j = (1:nre-1) / (nre);
j = [lcle lcle+(ucle-lcle)*j ucle];
e = j(1:nre)+diff(j)/2;
ep = [e(1:nre-1)+diff(e)/2 ucle];
em = [lcle e(2:nre)-diff(e)/2];
norme = diff (normcdf([em(1) ep]));
i = (1:nrx-1) / (nrx);
i = [lcl lcl+(ucl-lcl)*i ucl];
xp = i(2:nrx+1);
xm = i(1:nrx);
x = i(1:nrx)+diff(i)/2;
dx = x(2) - x(1);
o = (xp-xm)'*(ep-em);
g =  1/(2*pi*sqrt(1-corr^2)*sigmax) * exp( (-1/(2*(1-corr^2))) * ...
         ( (x'*ones(1,prod(size(e)))/sigmax).^2 ...
            - 2*corr*x'*e/sigmax + ...
           (ones(prod(size(x)),1)*e).^2));
```

```
p = o.*g;
arl(index,1)=1/(1-sum(sum(p)))



disp (sprintf ('phi = %6.3f  theta = %6.3f  lim = %6.3f
sigmax = %5.3f  rho = %5.3f',phi, theta, lim, sigmax, rho));

% compute the Markov transition matrix T
% where the columns in T correspond to
% (old x, old e, new x, new e)
  disp('Making Markov Transition Matrix')
  T = zeros( prod(size(find(p>=tiny)))*nre ,4);
  count = 0;
  for k = 1:nrx, for l = 1:nre,
     if (p(k,l) > tiny)
        xnew = phi*x(k) - theta*e(l) + e;
        kk = ceil ((xnew-lcl)/dx);
        for ll = 1: nre,
            if (kk(ll) > 0)
              if (kk(ll) > nrx)
                 break;
              else
                count = count+1;
                T(count,1:4)=[k l kk(ll) ll];
              end;
            end;
        end;
      end;
  end; end;
  elapse = etime(clock,start);
  disp(sprintf('Elapsed time (in min): %6.1f', elapse/60))

% Run thru the loops
disp('Starting Loops')
for loop = 1:nrloops,
  if (nrloops == 0) break; end;
% do the markov conversion
  pin = p / sum(sum(p));
  p = zeros (nrx, nre);
  for c = 1:count,
     p(T(c,3), T(c,4)) = p(T(c,3), T(c,4)) + pin(T(c,1),T(c,2));
  end;
  p = p .* (ones(nrx,1)*norme);
  temp = sum(sum(p));
  arl(index, loop+1) = 1/(1-temp);
  disp('ARL3'); disp(arl);
end; % loop loop

perc   = ((phiindex-1)*(nrlimpoints) + limindex) / (nrphipoints * nrlimpoints);
elapse = etime(clock,start);
left   = ((elapse/perc)-elapse)/60;
disp   (sprintf('%5.3f pct done %6.1f min left. (%6.1f min in)', ...
       perc, left, elapse/60))

end; % limindex loop
end; % phiindex loop
```

```
% display run time info
elapse = etime(clock,start);
disp(sprintf('Total time (in min): %6.1f', elapse/60))
```

Appendix C. Matlab Code for Numerically Approximating the Joint Probability Density Function for a Pure AR(1) Process.

```
start=clock;
phipoints = [0.95 0.475 0 -.475 -.95];
limpoints = [1.5 1.75 2 2.25 2.5 2.75 3 3.25];
theta   = 0 ;
nrx     = 1001;
nre     = 33;
nrloops = 31;

disp (sprintf (’nrx = %i  nre = %i  nrloops = %i’,nrx, nre, nrloops));
nrphipoints = size(phipoints,1)*size(phipoints,2);
nrlimpoints = size(limpoints,1)*size(limpoints,2);
indexsize = nrphipoints*nrlimpoints;
arl = zeros (indexsize, nrloops+1);
for phiindex = 1:nrphipoints,
for limindex = 1:nrlimpoints,
index = (phiindex-1)*nrlimpoints + limindex;
phi = phipoints(phiindex);
lim = limpoints(limindex);
phicode(index) = phi;
limcode(index) = lim;

gam0 = (1 + theta^2  - 2*phi*theta) / (1 - phi^2);
sigmax = sqrt(gam0);
rho = (1-theta*phi)*(phi-theta)/(1+theta^2-2*phi*theta);
```

```
lcl = -lim*sigmax;

ucl =  lim*sigmax;

lcle = -6;

ucle = 6;

dopt = 1 - (normcdf(ucl) - normcdf(lcl));

d2 = dopt / (dopt+1);


j = (1:nre-1) / (nre);

j = [lcle lcle+(ucle-lcle)*j ucle];

e = j(1:nre)+diff(j)/2;

ep = [e(1:nre-1)+diff(e)/2 ucle];

em = [lcle e(2:nre)-diff(e)/2];

norme = diff (normcdf([em(1) ep]));

i = (1:nrx-1) / (nrx);

i = [lcl lcl+(ucl-lcl)*i ucl];

xp = i(2:nrx+1);

xm = i(1:nrx);

x = i(1:nrx)+diff(i)/2;

dx = x(2) - x(1);

g =  1/(sqrt(2*pi)*sigmax) * exp( (-1/2) *(x/sigmax).^2);

p = dx.*g;

arl(index,1)=1/(1-sum(p));



disp (sprintf ('phi = %6.3f  theta = %6.3f  lim = %6.3f

sigmax = %5.3f  rho = %5.3f',phi, theta, lim, sigmax, rho));


% compute the Markov transition matrix T
% where the columns in T correspond to
% (old x, old e, new x, new e)
  disp('Making Markov Transition Matrix')
  T = zeros( nrx, nrx);
  for k = 1:nrx,
     T(:,k) = dx*normpdf(x(k)-phi*x');
  end;
  elapse = etime(clock,start);
  disp(sprintf('Elapsed time (in min): %6.1f', elapse/60))

% Run thru the loops
disp('Starting Loops')
for loop = 1:nrloops,
  if (nrloops == 0) break; end;
% do the markov conversion
  pin = p / sum(p);
  p = pin*T;
  temp = sum(p);
  arl(index, loop+1) = 1/(1-temp);
%  disp('ARL'); disp(arl);
end; % loop loop
%  disp('ARL'); disp(arl);

perc   = ((phiindex-1)*(nrlimpoints) + limindex) / (nrphipoints * nrlimpoints);
elapse = etime(clock,start);
left   = ((elapse/perc)-elapse)/60;
disp   (sprintf('%5.3f pct done %6.1f min left. (%6.1f min in)', ...
       perc, left, elapse/60))
```

```
end; % limindex loop
end; % phiindex loop

disp('ARL'); disp(arl);
disp ([phicode' limcode' arl(:,1) arl(:,nrloops+1)])
% display run time info
elapse = etime(clock,start);
disp(sprintf('Total time (in min): %6.1f', elapse/60))
```

Appendix D. Matlab Code for Implementing and Testing the Method for Monitoring Process Capability.

```
% Set the parameters for this run
designpt = 1;
nrruns     = 200;
acsize     = 0;

maxobs     = 5000;
stime      = cputime;
window     = 30;
cpktrig    = .90;
SCCtrig    = 1/370;
maxphi     = .97;
maxtheta   = 2;
lambda     = .15;

% load design point info
load -ascii datalime
load -ascii dataphitheta
% Initialize arrays and variables
phi    = dataphitheta(designpt,1);
theta  = dataphitheta(designpt,2);
Le     = datalime(designpt,:);
tau    = 0;
gam0   = (1 + theta*theta  - 2*phi*theta) / (1 - phi*phi);
sigmax = sqrt(gam0);
```

```
usl =  4*sigmax;   ucl =  Le(2);   upl = usl;
lsl = -4*sigmax;   lcl = -Le(2);   lpl = lsl;
truespace = [phi theta 0 0];
signal    = zeros (nrruns, 2);
for i = 1:8,
   Zi(i) = phi^(i-1)*(phi-theta);
   Smult(i) = sqrt(1 + sum(Zi(1:i-1).^2));
end;

% set up the smoothing variables

disp (sprintf ('designpt %i acsize %4.1f maxobs %i',designpt,acsize,maxobs));

for run = 1:nrruns,

% generate an initially controlled time series
   x= inf;
   while (max(x) > usl | min(x) < lsl),
       [x e] = armagen(truespace, window);
   end;
   xac = x;
   % set the starting values for smoothing variables
   th = zarmax (xac-mean(xac), [1 1]);
   smphi    = -th(3,1)*.8;
   smtheta = -th(3,2)*.9;
   smse     = sqrt(th(1,1))*.9;
   smxi = 0;
```

```
% loop until a signal is generated
obs = 0;
while (signal(run,1) == 0 & signal(run,2) == 0),

% extend the time-series one observation
obs = obs+1;
if (obs > maxobs),
   signal(run,1) = maxobs+1; disp('EXCEEDED MAXOBS'); break;
end;
[xout eout] = armaextend (truespace, x(window), e(window), 1);
x   = [x(2:window); xout(1)];
e   = [e(2:window); eout(1)];
xac = [xac(2:window); xout(1)+acsize*sigmax];

% fit an ARMA(1,1) model
meanxac = mean(xac);
th = zarmax (xac-meanxac, [1 1]);
fitse    = sqrt(th(1,1));
fittheta = -th(3,2);
fitphi   = -th(3,1);
fitxi    = meanxac * (1-fitphi);
if (fittheta > maxtheta) fittheta= maxtheta; end;
if (fittheta <-maxtheta) fittheta=-maxtheta; end;
if (fitphi > maxphi) fitphi= maxphi; end;
if (fitphi <-maxphi) fitphi=-maxphi; end;

fitspace = [fitphi fittheta fitxi 0];
fite1 = fmin ('armaerre', -3, 3, [], fitspace, xac);
```

```
fitspace(4) = fite1;

% compute Cpk and SWalpha
smphi    = lambda*fitphi   + (1-lambda)*smphi;
smtheta  = lambda*fittheta + (1-lambda)*smtheta;
smse     = lambda*fitse    + (1-lambda)*smse;
smxi     = lambda*fitxi    + (1-lambda)*smxi;
smsigmax = sqrt( (1+smtheta.^2-2*smtheta.*smphi)./(1-smphi.^2) );
smmean = smxi / (1-smphi);
smcpk = min(usl-smmean, smmean-lsl)./(3*smsigmax*smse);

% test for lack of capability
if (smcpk < cpktrig)
   signal(run,1) = obs;
end;

% FIR code
if (signal(run,2) == 0),
   [xhat, ehat]=armafit (fitspace, xac);
   swalpha = sw30(ehat);
   if (swalpha < SCCtrig)
       disp(sprintf('SCC ALARM at obs+%i   1/swalpha = %6.1f' ...
             ,obs,1/swalpha));
       signal(run,2) = obs;
    end;
 end;
 % End of FIR code
```

```
   end;

end;

disp (sprintf (’designpt %i acsize %4.1f maxobs %i’,designpt,acsize,maxobs));
etime = cputime;
disp(sprintf(’Elapsed time (in min) %7.1f For %i runs’, (etime-stime)/60, nrruns));
rlsig = max (signal’);
arl = mean(rlsig);
srl = std(rlsig);
disp (sprintf (’ARL %8.1f   SRL %8.1f’,arl,srl));
disp (sprintf (’pct FAC %4.1f   pct FAN %4.1f’, ...
     100*sum(signal(:,1)>0)/nrruns, 100*sum(signal(:,2)>0)/nrruns));
disp (sprintf (’90 pct CI on ARL: (%8.1f - %8.1f)’, ...
     arl-1.645*srl/sqrt(nrruns), arl+1.645*srl/sqrt(nrruns) ) );
```

Appendix E. Copy of Key Results (Wardell, Moskowitz and Plante, 1994).

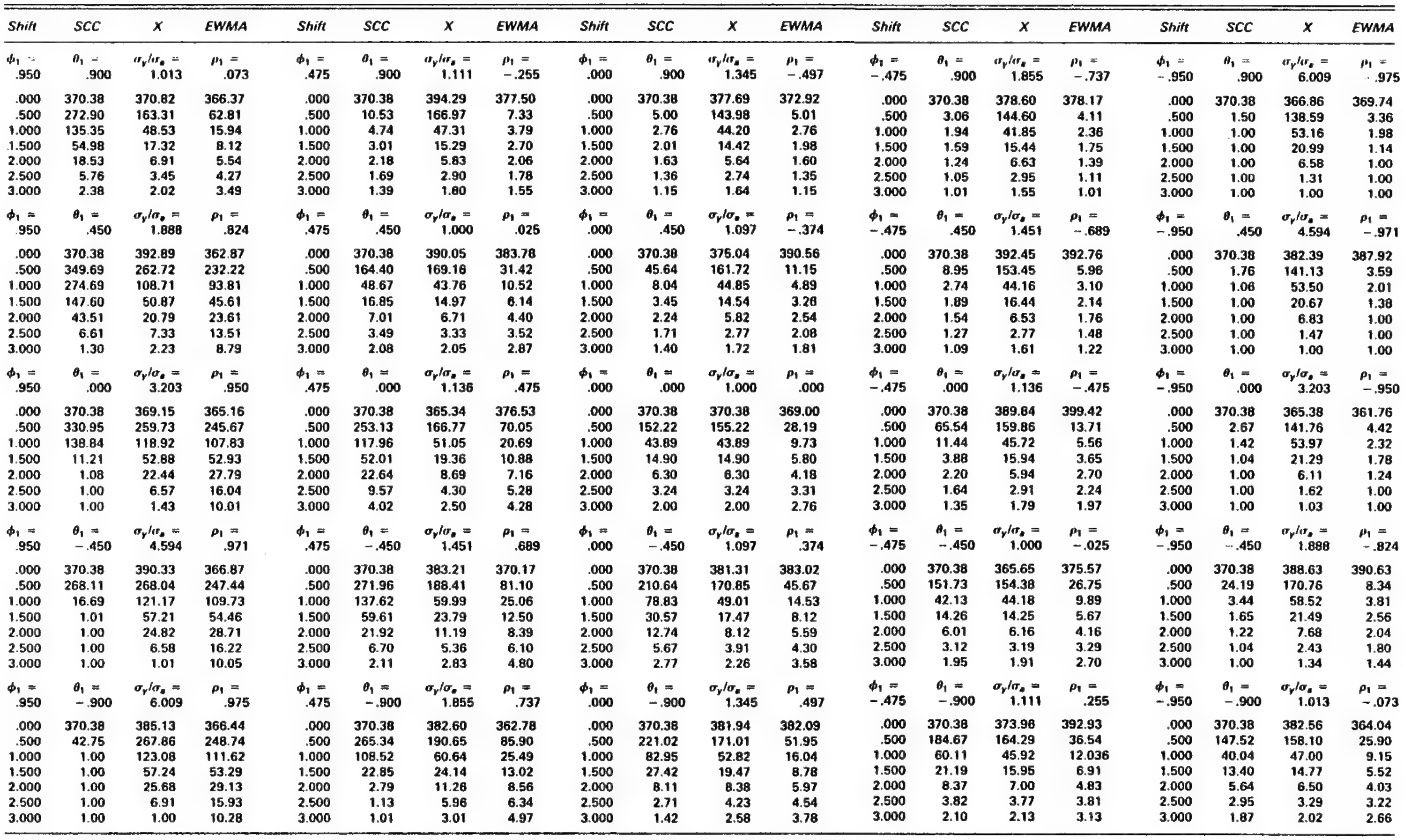


Table 4. Comparison of ARL Values of the Special-Cause Chart (SCC), the Shewhart (X) Chart, and the Exponentially Weighted Moving Average (EWMA) Chart for Various ARMA(1,1) Parameters

Shift	SCC	X	EWMA	Shift	SCC	X	EWMA	Shift	SCC	X	EWMA	Shift	SCC	X	EWMA	Shift	SCC	X	EWMA
$\phi_1 =$ .950	$\theta_1 =$ .900	$\sigma_y/\sigma_e =$ 1.013	$\rho_1 =$ .073	$\phi_1 =$ .475	$\theta_1 =$ .900	$\sigma_y/\sigma_e =$ 1.111	$\rho_1 =$ −.255	$\phi_1 =$ .000	$\theta_1 =$ .900	$\sigma_y/\sigma_e =$ 1.345	$\rho_1 =$ −.497	$\phi_1 =$ −.475	$\theta_1 =$ .900	$\sigma_y/\sigma_e =$ 1.855	$\rho_1 =$ −.737	$\phi_1 =$ −.950	$\theta_1 =$ .900	$\sigma_y/\sigma_e =$ 6.009	$\rho_1 =$ −.975
.000	370.38	370.82	366.37	.000	370.38	394.29	377.50	.000	370.38	377.69	372.92	.000	370.38	378.60	378.17	.000	370.38	366.86	369.74
.500	272.90	163.31	62.81	.500	10.53	166.97	7.33	.500	5.00	143.98	5.01	.500	3.06	144.60	4.11	.500	1.50	138.59	3.36
1.000	135.35	48.53	15.94	1.000	4.74	47.31	3.79	1.000	2.76	44.20	2.76	1.000	1.94	41.85	2.36	1.000	1.00	53.16	1.98
1.500	54.98	17.32	8.12	1.500	3.01	15.29	2.70	1.500	2.01	14.42	1.98	1.500	1.59	15.44	1.75	1.500	1.00	20.99	1.14
2.000	18.53	6.91	5.54	2.000	2.18	5.83	2.06	2.000	1.63	5.64	1.60	2.000	1.24	6.63	1.39	2.000	1.00	6.58	1.00
2.500	5.76	3.45	4.27	2.500	1.69	2.90	1.78	2.500	1.36	2.74	1.35	2.500	1.05	2.95	1.11	2.500	1.00	1.31	1.00
3.000	2.38	2.02	3.49	3.000	1.39	1.80	1.55	3.000	1.15	1.64	1.15	3.000	1.01	1.55	1.01	3.000	1.00	1.00	1.00
$\phi_1 =$ .950	$\theta_1 =$ .450	$\sigma_y/\sigma_e =$ 1.888	$\rho_1 =$ .824	$\phi_1 =$ .475	$\theta_1 =$ .450	$\sigma_y/\sigma_e =$ 1.000	$\rho_1 =$ .025	$\phi_1 =$ .000	$\theta_1 =$ .450	$\sigma_y/\sigma_e =$ 1.097	$\rho_1 =$ −.374	$\phi_1 =$ −.475	$\theta_1 =$ .450	$\sigma_y/\sigma_e =$ 1.451	$\rho_1 =$ −.689	$\phi_1 =$ −.950	$\theta_1 =$ .450	$\sigma_y/\sigma_e =$ 4.594	$\rho_1 =$ −.971
.000	370.38	392.89	362.87	.000	370.38	390.05	383.78	.000	370.38	375.04	390.56	.000	370.38	392.45	392.76	.000	370.38	382.39	387.92
.500	349.69	262.72	232.22	.500	164.40	169.16	31.42	.500	45.64	161.72	11.15	.500	8.95	153.45	5.96	.500	1.76	141.13	3.59
1.000	274.69	108.71	93.81	1.000	48.67	43.76	10.52	1.000	8.04	44.85	4.89	1.000	2.74	44.16	3.10	1.000	1.06	53.50	2.01
1.500	147.60	50.87	45.61	1.500	16.85	14.97	6.14	1.500	3.45	14.54	3.26	1.500	1.89	16.44	2.14	1.500	1.00	20.67	1.38
2.000	43.51	20.79	23.61	2.000	7.01	6.71	4.40	2.000	2.24	5.82	2.54	2.000	1.54	6.53	1.76	2.000	1.00	6.83	1.00
2.500	6.61	7.33	13.51	2.500	3.49	3.33	3.52	2.500	1.71	2.77	2.08	2.500	1.27	2.77	1.48	2.500	1.00	1.47	1.00
3.000	1.30	2.23	8.79	3.000	2.08	2.05	2.87	3.000	1.40	1.72	1.81	3.000	1.09	1.61	1.22	3.000	1.00	1.00	1.00
$\phi_1 =$ .950	$\theta_1 =$ .000	$\sigma_y/\sigma_e =$ 3.203	$\rho_1 =$ .950	$\phi_1 =$ .475	$\theta_1 =$ .000	$\sigma_y/\sigma_e =$ 1.136	$\rho_1 =$ .475	$\phi_1 =$ .000	$\theta_1 =$ .000	$\sigma_y/\sigma_e =$ 1.000	$\rho_1 =$ .000	$\phi_1 =$ −.475	$\theta_1 =$ .000	$\sigma_y/\sigma_e =$ 1.136	$\rho_1 =$ −.475	$\phi_1 =$ −.950	$\theta_1 =$ .000	$\sigma_y/\sigma_e =$ 3.203	$\rho_1 =$ −.950
.000	370.38	369.15	365.16	.000	370.38	365.34	376.53	.000	370.38	370.38	369.00	.000	370.38	389.84	399.42	.000	370.38	365.38	361.76
.500	330.95	259.73	245.67	.500	253.13	166.77	70.05	.500	152.22	155.22	28.19	.500	65.54	159.86	13.71	.500	2.67	141.76	4.42
1.000	138.84	118.92	107.83	1.000	117.96	51.05	20.69	1.000	43.89	43.89	9.73	1.000	11.44	45.72	5.56	1.000	1.42	53.97	2.32
1.500	11.21	52.88	52.93	1.500	52.01	19.36	10.88	1.500	14.90	14.90	5.80	1.500	3.88	15.94	3.65	1.500	1.04	21.29	1.78
2.000	1.08	22.44	27.79	2.000	22.64	8.69	7.16	2.000	6.30	6.30	4.18	2.000	2.20	5.94	2.70	2.000	1.00	6.11	1.24
2.500	1.00	6.57	16.04	2.500	9.57	4.30	5.28	2.500	3.24	3.24	3.31	2.500	1.64	2.91	2.24	2.500	1.00	1.62	1.00
3.000	1.00	1.43	10.01	3.000	4.02	2.50	4.28	3.000	2.00	2.00	2.76	3.000	1.35	1.79	1.97	3.000	1.00	1.03	1.00
$\phi_1 =$ .950	$\theta_1 =$ −.450	$\sigma_y/\sigma_e =$ 4.594	$\rho_1 =$ .971	$\phi_1 =$ .475	$\theta_1 =$ −.450	$\sigma_y/\sigma_e =$ 1.451	$\rho_1 =$ .689	$\phi_1 =$ .000	$\theta_1 =$ −.450	$\sigma_y/\sigma_e =$ 1.097	$\rho_1 =$ .374	$\phi_1 =$ −.475	$\theta_1 =$ −.450	$\sigma_y/\sigma_e =$ 1.000	$\rho_1 =$ −.025	$\phi_1 =$ −.950	$\theta_1 =$ −.450	$\sigma_y/\sigma_e =$ 1.888	$\rho_1 =$ −.824
.000	370.38	390.33	366.87	.000	370.38	383.21	370.17	.000	370.38	381.31	383.02	.000	370.38	365.65	375.57	.000	370.38	388.63	390.63
.500	268.11	268.04	247.44	.500	271.96	188.41	81.10	.500	210.64	170.85	45.67	.500	151.73	154.38	26.75	.500	24.19	170.76	8.34
1.000	16.69	121.17	109.73	1.000	137.62	59.99	25.06	1.000	78.83	49.01	14.53	1.000	42.13	44.18	9.89	1.000	3.44	58.52	3.81
1.500	1.01	57.21	54.46	1.500	59.61	23.79	12.50	1.500	30.57	17.47	8.12	1.500	14.26	14.25	5.67	1.500	1.65	21.49	2.56
2.000	1.00	24.82	28.71	2.000	21.92	11.19	8.39	2.000	12.74	8.12	5.59	2.000	6.01	6.16	4.16	2.000	1.22	7.68	2.04
2.500	1.00	6.58	16.22	2.500	6.70	5.36	6.10	2.500	5.67	3.91	4.30	2.500	3.12	3.19	3.29	2.500	1.04	2.43	1.80
3.000	1.00	1.01	10.05	3.000	2.11	2.83	4.80	3.000	2.77	2.26	3.58	3.000	1.95	1.91	2.70	3.000	1.00	1.34	1.44
$\phi_1 =$ .950	$\theta_1 =$ −.900	$\sigma_y/\sigma_e =$ 6.009	$\rho_1 =$ .975	$\phi_1 =$ .475	$\theta_1 =$ −.900	$\sigma_y/\sigma_e =$ 1.855	$\rho_1 =$ .737	$\phi_1 =$ .000	$\theta_1 =$ −.900	$\sigma_y/\sigma_e =$ 1.345	$\rho_1 =$ .497	$\phi_1 =$ −.475	$\theta_1 =$ −.900	$\sigma_y/\sigma_e =$ 1.111	$\rho_1 =$ .255	$\phi_1 =$ −.950	$\theta_1 =$ −.900	$\sigma_y/\sigma_e =$ 1.013	$\rho_1 =$ −.073
.000	370.38	385.13	366.44	.000	370.38	382.60	362.78	.000	370.38	381.94	382.09	.000	370.38	373.98	392.93	.000	370.38	382.56	364.04
.500	42.75	267.86	248.74	.500	265.34	190.65	85.90	.500	221.02	171.01	51.95	.500	184.67	164.29	36.54	.500	147.52	158.10	25.90
1.000	1.00	123.08	111.62	1.000	108.52	60.64	25.49	1.000	82.95	52.82	16.04	1.000	60.11	45.92	12.036	1.000	40.04	47.00	9.15
1.500	1.00	57.24	53.29	1.500	22.85	24.14	13.02	1.500	27.42	19.47	8.78	1.500	21.19	15.95	6.91	1.500	13.40	14.77	5.52
2.000	1.00	25.68	29.13	2.000	2.79	11.26	8.56	2.000	8.11	8.38	5.97	2.000	8.37	7.00	4.83	2.000	5.64	6.50	4.03
2.500	1.00	6.91	15.93	2.500	1.13	5.96	6.34	2.500	2.71	4.23	4.54	2.500	3.82	3.77	3.81	2.500	2.95	3.29	3.22
3.000	1.00	1.00	10.28	3.000	1.01	3.01	4.97	3.000	1.42	2.58	3.78	3.000	2.10	2.13	3.13	3.000	1.87	2.02	2.66

NOTE: ϕ_1 = autoregressive parameter; θ_1 = moving average parameter; σ_y = standard deviation of the observations; σ_e = standard deviation of the random error; ρ_1 = first order autocorrelation coefficient; SCC = special-cause chart; y = individuals (Shewhart) chart; EWMA = exponentially weighted moving average chart.

Bibliography

Alwan, L. C. and Roberts, H. V. (1988). Time-series modeling for statistical process control. *Journal of Business & Economic Statistics*, 6(1):87–95.

American Society for Quality Control, Statistics Division (1983). *Glossary and tables for statistical quality control.* ASQC Quality Press, Milwaukee, Second edition.

Aroian, L. A. and Levene, H. (1950). The effectiveness of quality control charts. *Journal of the American Statistical Association*, 45:520–529.

Bagshaw, M. and Johnson, R. A. (1977). Sequential procedures for detecting parameter changes in a time-series model. *Journal of the American Statistical Association*, 72(359):593–597.

Beauregard, M. R., Mikulak, R. J., and Olson, B. A. (1992). *A Practical Guide to Statistical Quality Improvement.* Van Nostrand Reinhold, New York.

Berthouex, P. M., Hunter, W. G., and Pallsen, L. (1978). Monitoring sewage treatment plants: some quality control aspects. *Journal of Quality Technology*, 10(4):139–149.

Bissell, A. F. (1990). How reliable is your capability index? *Applied Statistics*, 39(3):331–340.

Box, G. and Kramer, T. (1992). Statistical process monitoring and feedback adjustment-a discussion. *Technometrics*, 34(3):251–285.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis: forecasting and control.* Holden-Day Inc., San Francisco.

Box, G. E. P. and Tiao, G. C. (1975). Intervention analysis with applications to economic and environmental problems. *Journal of the American Statistical Association*, 70(349):70–79.

Boyles, R. A. (1991). The Taguchi capability index. *Journal of Quality Technology*, 23(1):17–26.

Champ, C. W. and Woodall, W. H. (1987). Exact results for Shewart control charts with supplementary runs rules. *Technometrics*, 29(4):393–399.

Chan, L. K., Cheng, S. W., and Spiring, F. A. (1988). A new measure of process capability: C_{pm}. *Journal of Quality Technology*, 20(3):162–175.

Cheng, S. W. (1994). Practical implementation of the process capability indices. *Quality Engineering*, 7(2):239–259.

Chung, K.-J. (1992). Economically optimal determination of the parameters of CUSUM charts. *International Journal of Quality and Reliability Management*, 9(6):8–17.

Conover, W. J. (1980). *Practical Nonparametric Statistics.* John Wiley & Sons, New York, Second edition.

DeGroot, M. H. (1989). *Probability and Statistics.* Addison-Wesley Publishing Company, Inc., Reading, Massachusetts, Second edition.

Doty, L. A. (1990). *Statistical Process Control.* Industrial Press Inc., New York, First edition.

Ermer, D. S., Chow, M. C., and Wu, S. M. (1979). A time series control chart for a nuclear reactor. *Proceedings of the Annual Reliability and Maintainability Symposium*, pages 92–97.

Franklin, L. A. and Wasserman, G. S. (1992). Bootstrap lower confidence limits for capability indices. *Journal of Quality Technology*, 24(4):196–210.

Goel, A. L. and Wu, S. M. (1973). Economically optimal design of CUSUM charts. *Management Science*, 19(11):1271–1282.

Hogg, R. V. and Craig, A. T. (1978). *Introduction to Mathematical Statistics.* Macmillan Publishing Co., Inc., New York, Fourth edition.

Igelwicz, B. and Hoaglin, D. C. (1993). *How to detect and handle outliers.* American Society for Quality Control, Milwaukee. Volume 16 of the ASQC Basic References in Quality Control: Statistical Techniques.

Kane, V. E. (1986a). Corrigenda: Process capability indices. *Journal of Quality Technology*, 18(4):265.

Kane, V. E. (1986b). Process capability indices. *Journal of Quality Technology*, 18(1):41–52.

Ljung, L. (1992). *System Identification Toolbox for use with Matlab.* The MathWorks, Inc., Natick, Mass.

Long, J. M. and DeCoste, M. J. (1988). Capability studies involving tool wear. *ASQC Quality Conference Transactions*, pages 590–596.

Lorenzen, T. J. and Vance, L. C. (1986). The economic design of control charts: a unified approach. *Technometrics*, 28(1):3–10.

Lucas, J. M. and Crosier, R. B. (1982a). Fast initial response for CUSUM quality-control schemes: give your CUSUM a head start. *Technometrics*, 24(3):199–205.

Lucas, J. M. and Crosier, R. B. (1982b). Robust CUSUM: a robustness study for CUSUM quality control schemes. *Communications in Statistics- Theory and Methods*, 11(23):2669–2687.

Lucas, J. M. and Sacucci, M. S. (1990). Exponentially weighted moving average control schemes: properties and enhancements. *Technometrics*, 32(1):1–29.

MacGregor, J. F. (1988). On-line statistical process control. *Chemical Engineering Progress*, pages 21–31.

MacGregor, J. F. and Harris, T. J. (1990). Discussion of EWMA control schemes by Lucas and Saccucci. *Technometrics*, 32(1):23–26.

MacGregor, J. F. and Harris, T. J. (1993). The exponentially weighted moving variance. *Journal of Quality Technology*, 25(2).

Maragah, H. D. and Woodall, W. H. (1992). The effect of autocorrelation on the retrospective X-chart. *Journal of Statistical Computational Simulation*, 40:29–42.

Marcucci, M. O. and Beazley, C. C. (1988). Capability indices: Process performance measures. *ASQC Quality Conference Transactions*, pages 516–523.

Montgomery, D. C. (1991). *Introduction to Statistical Quality Control.* SPIE AIPR Workshop, Second edition.

Montgomery, D. C. and Friedman, D. J. (1989). Statistical process control in a computer-integrated manufacturing environment. *Statistical Process Control in Automated Manufacturing.* edited by J. Bert Keats.

Montgomery, D. C., Johnson, L. A., and Gardner, J. S. (1990). *Forecasting and Time Series Analysis.* McGraw-Hill, Inc., Second edition.

Montgomery, D. C., Keats, J. B., Runger, G. C., and Messina, W. S. (1994). Integrating statistical process control and engineering process control. *Journal of Quality Technology*, 26(2):79–87.

Montgomery, D. C. and Mastrangelo, C. M. (1991). Some statistical process control methods for autocorrelated data. *Journal of Quality Technology*, 23(3).

Page, E. S. (1954). Continuous inspection schemes. *Biometrika*, 41:100–114.

Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika*, 42:523–527.

Pritsker, A. A. B. (1986). *Introduction to Simulation and SLAM II.* John Wiley & Sons, Inc., New York, Third edition.

Quesenberry, C. P. (1988). An SPC approach to compensating a tool-wear process. *Journal of Quality Technology*, 20(4):220–229.

Roberts, S. W. (1959). Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250.

Schilling, E. D. and Nelson, P. R. (1976). The effect of non-normality on the control limits of $\bar{X}$ charts. *Journal of Quality Technology*, 8(4):183–188.

Shapiro, S. S. (1980). *How to test normality and other distributional assumptions.* American Society for Quality Control, Milwaukee. Volume 3 of the ASQC Basic References in Quality Control: Statistical Techniques.

Shewhart, W. A. (1931). *Economic Control of Quality of Manufactured Product.* D. Van Nostrand Company, Inc., New York. 50th Anniversary Commemorative Reissue.

Shewhart, W. A. (1986). *Statistical Method from the Viewpoint of Quality Control.* General Publishing Company, Ltd., Toronto. Republication of work originally published in 1939 with original foreword by W. Edwards Deming.

Snee, R. D. (1990). Statistical thinking and its contribution to total quality. *The American Statistician*, 44(2):116–121.

Taguchi, G. (1985). Quality engineering in Japan. *Communications in Statistics- Theory and Methods*, 14(11):2785–2801.

Taguchi, G. and Wu, Y. (1980). *Introduction to Off-line Quality Control.* Central Japan Quality Control Association, Nagoya, Japan.

Tukey, J. W. (1961). Discussion, emphasizing the connection between analysis of variance and spectrum analysis. *Technometrics*, 3(2):191–219.

Vander Wiel, S. A., Tucker, W. T., Faltin, F. W., and Doganaksoy, N. (1992). Algorithmic statistical process control: Concepts and an application. *Technometrics*, 34(3):286–297.

Wardell, D. G., Moskowitz, H., and Plante, R. D. (1992). Control charts in the presence of data correlation. *Management Science*, 38(8):1084–1105.

Wardell, D. G., Moskowitz, H., and Plante, R. D. (1994). Run-length distributions of special-cause control charts for correlated processes. *Technometrics*, 36(1):3–17.

Wheeler, D. J. and Chambers, D. S. (1992). *Understanding Statistical Process Control.* SPC Press, Inc., Knoxville, Tennessee, Second edition.

Yourstone, S. E. and Montgomery, D. C. (1989). A time-series approach to discrete real-time process quality control. *Quality and Reliability Engineering International*, 5:309–317.

Vita

Major Daniel J. Zalewski was born on 8 September 1960 in Detroit, Michigan. In 1978, he graduated from Cherry Hill High School and was accepted at the United States Air Force Academy. In 1983, he graduated from the Academy with a Bachelor of Science Degree in Computer Technology and Operations Research. His first assignment was with the Air Force Data Services Center at the Pentagon in Washington, D.C. Following that assignment, the Air Force sponsored him for a Master's Degree at George Mason University in Fairfax, Virginia. He received his Master of Science Degree in 1988. His subsequent assignment was in the Intelligence Database Support Branch at Strategic Air Command, Offut Air Force Base, Nebraska. He entered the doctoral program in Operations Research in 1992 at the Air Force Institute of Technology.

Permanent address: 29679 Steinhauer
Inkster, Michigan 48141

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
August 1995

3. REPORT TYPE AND DATES COVERED
Doctoral Dissertation

4. TITLE AND SUBTITLE
METHODS FOR MONITORING PROCESS CONTROL AND CAPABILITY IN THE PRESENCE OF AUTOCORRELATION

5. FUNDING NUMBERS

6. AUTHOR(S)
Daniel J. Zalewski, Major, USAF

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Air Force Institute of Technology, WPAFB OH 45433-7765

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT/DS/ENS/95-02

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
ASC/YCD (Jim Arnold)
2600 Paramount Place
Fairborn, OH 45324-6766

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

When standard control charts are applied to a process whose measurements of quality exhibit autocorrelation, the performance of those charts can be considerably different than that expected when no autocorrelation is present. To model this performance, the existing definitions of assignable and chance causes of variation are extended to account for the variation induced by the autocorrelation structure. The application of statistical thinking toward continuous process improvement using the proposed taxonomy is discussed. A method to select control limits which yield a specified average run length in the absence of assignable causes of variation and which is suitable for use on processes whose behavior can be modelled as an ARMA(1,1) process is developed. The current paradigm for process improvement is centered around monitoring the state of statistical control. A new paradigm, based upon monitoring process capability, is proposed. The time-varying aspects of capability are highlighted. A capability monitoring system for stationary ARMA(1,1) processes is developed and compared to other standard methods. The benefits of additional knowledge are demonstrated by simulating the response of capability monitoring systems tailored to independent normal and mixed ARMA(1,1) models to shifts in the mean and variance.

14. SUBJECT TERMS
Quality Improvement, Statistical Process Control, Capability, ARMA(1,1)

15. NUMBER OF PAGES
170

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL